PORTUÑOL TRANSLATIONS

460 PIERCE STREET
MONTEREY, CA 93940

Team Members: Nathalia Rio Preto,
Isabel González-Gutiérrez, and Fiona Maloney-McCrystle

PROJECT PROPOSAL

PREPARED ON 2/19/20 FOR:
Professor Adam Wooten
460 Pierce St.
Monterey, CA 93940

PROJECT OVERVIEW:

This pilot project seeks to estimate the amount of time and costs implied in training a statistical machine translation engine designed to translate ocean-related conservation reports. In order to arrive at these estimates, our pilot project will contain the following broad steps: identification and preparation of data sets, rounds of testing with adjusted data as needed, and the analysis of efficiency, cost, and quality metrics, to be compared at multiple points during the process. At the end of this month-long pilot, we will deliver an updated estimate of time and costs needed to complete the training of the described machine translation engine, as well as recommendations regarding that future training.

METRICS:

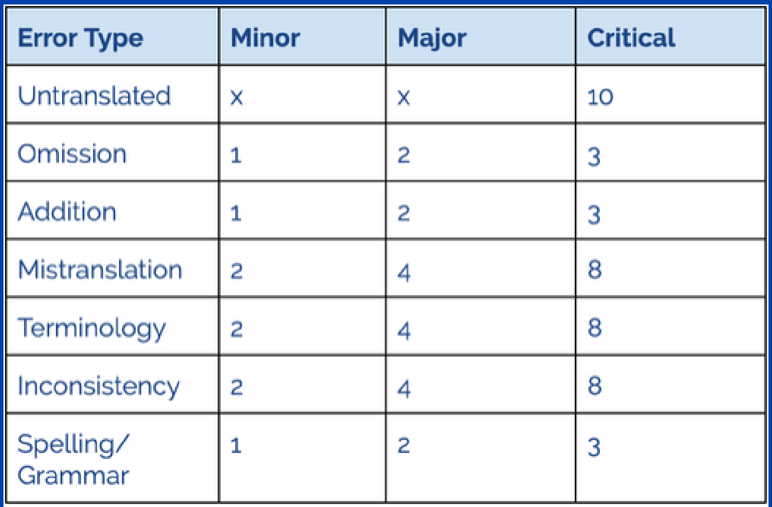
Progress will be measured and tracked through the following metrics:

-**Efficiency: The goal is to have human post-editing of machine translated content (PEMT) be 35% faster than human translation of that content.** This will be tracked using time comparisons between estimated time for human translation and the average of the time it takes three human translators to post-edit to acceptable quality (see metrics below). Comparisons will also be made between post-editing time required near the beginning and the end of the training window.

-**Cost: The goal is to have PEMT of content be 35% cheaper than human translation of that content.** This will also be monitored using the same time data collection described above, multiplied by our rate of $40/hr and compared to estimated human translation time and tracked from the beginning to the end of the project.

-**Quality: The goal is to have the translations produced by the MT engine achieve an acceptable score based on the quality metrics detailed below.** After two rounds of training, and again after the final round of training, human translators will post-edit, reviewing the document in accordance with the following error types and severities to ensure that it passes the threshold of acceptability:

Quality metrics have been adapted from error types on the MQM Scorecard, with weights and severities assigned by us.



Error Type	Minor	Major	Critical
Untranslated	x	x	10
Omission	1	2	3
Addition	1	2	3
Mistranslation	2	4	8
Terminology	2	4	8
Inconsistency	2	4	8
Spelling/ Grammar	1	2	3

If a text reaches 10 points per 500 words, it will not pass as acceptable.

DATA:

-**Training Data:** Annual reports and research reports from general conservation organizations and UN environmental programs

-**Tuning Data:** Ocean-specific reports from ocean conservation organizations (such as the Monterey Bay Aquarium) and the UN

-**Testing Data:** Report from the UN detailing progress towards SDG 14, which deals with conserving and sustainably using the oceans, seas, and marine resources

PROCESS:

Preparation Phase

- Determine scope and audience of the MT engine to be developed
- Gather the appropriate domain-specific data: any parallel texts, monolingual texts, or glossaries needed
- Convert PDFs to Word and align in SDL Trados to create TMX files
- Analyze and finalize data to ensure preliminary relevance and usefulness

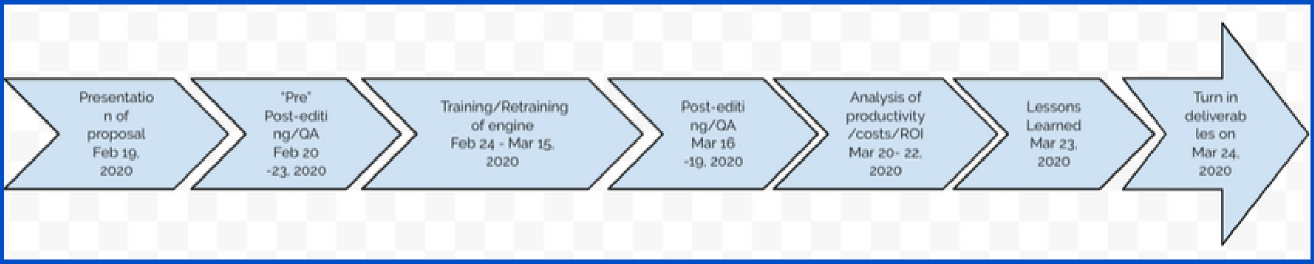
Production Phase

- Upload parallel texts to engine
- Conduct training cycles with Microsoft Custom Translator, carrying out as many rounds as needed (estimated between 10 and 15)
- Evaluate BLEU score produced by Custom Translator after each training cycle and adjust data accordingly (i.e. adding better human translations, adding more training data, adding more terminology-specific training data, eliminating poor quality training data, creating and adding a dictionary for specific terms, etc.)

Analysis + Finalization Phase

- At two points during project, perform a timed post edit using three human translators and average the times
- At each of these points, run a comparison analysis on cost, quality, and efficiency: PEMT vs human translator
- Use quality assurance metric specified above to ensure that translation reaches threshold of acceptability
- Create new project proposal with updated estimates, recommendations, and reflections

TIMELINE:



Presentation of proposal Feb 19, 2020

"Pre" Post-editing/QA Feb 20 -23, 2020

Training/Retraining of engine Feb 24 - Mar 15, 2020

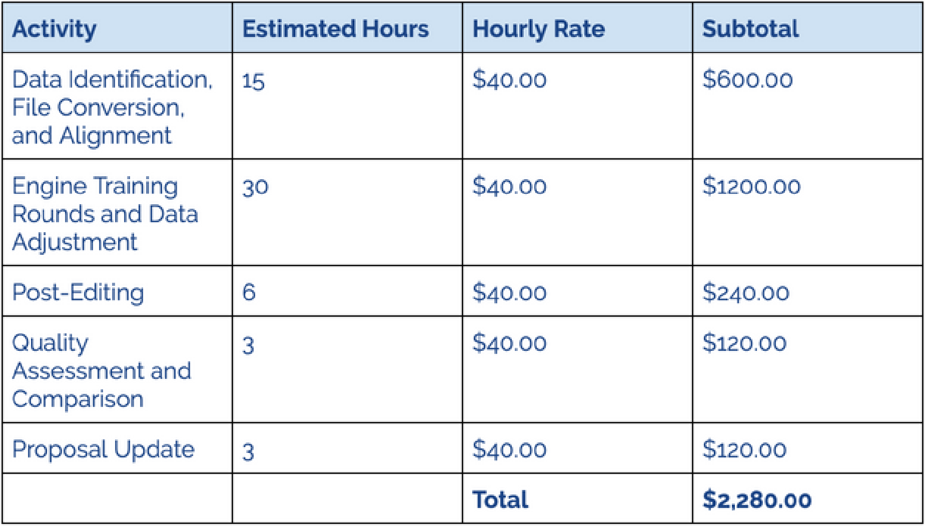
Post-editing/QA Mar 16 -19, 2020

Analysis of productivity /costs/ROI Mar 20- 22, 2020

Lessons Learned Mar 23, 2020

Turn in deliverables on Mar 24, 2020

ESTIMATED COSTS:



Activity	Estimated Hours	Hourly Rate	Subtotal
Data Identification, File Conversion, and Alignment	15	$40.00	$600.00
Engine Training Rounds and Data Adjustment	30	$40.00	$1200.00
Post-Editing	6	$40.00	$240.00
Quality Assessment and Comparison	3	$40.00	$120.00
Proposal Update	3	$40.00	$120.00
		Total	**$2,280.00**

DELIVERABLES:

At the end of the pilot project, the client will receive the following deliverables:

- Updated proposal for MT engine training project, including recommendations for future training of engine and updated efficiency, cost, and quality estimates
- Information about data sets used for training, tuning, and testing
- Bleu scores attained for each training cycle
- Log of changes made to data after each training cycle
- Glossary for specific domain, if created during training
- Cost-benefit analysis