MT for L10n:
How we build and evaluate MT systems at eBay

March 2017

Jose Luis Bonilla Sánchez - MTLS Manager

Contributors:
Silvio Picinini (MTLS team)
Kantan team

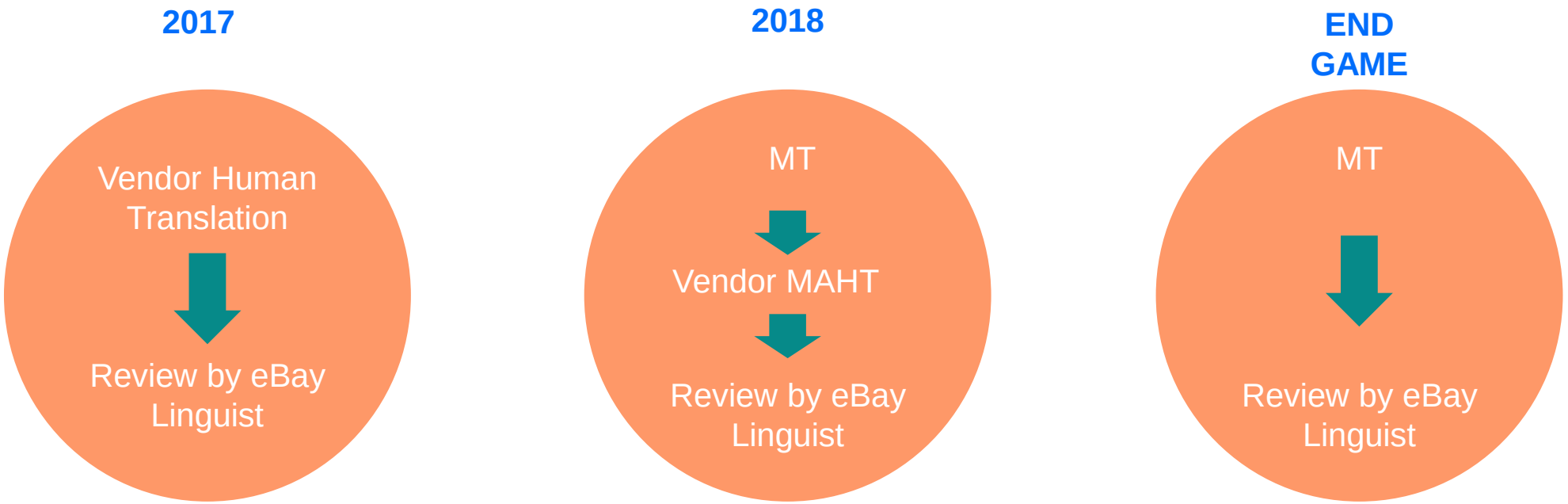
MT for L10n: How we build and evaluate MT systems at eBay

Agenda

- The L10n Roadmap
- The Master Pilot
- Phase I: Engine Building & Report-based Evaluation
- Phase II: Human Evaluation
- Conclusions

The eBay L10n Roadmap

L10n Roadmap: MT for All eBay-created content (Help, UI, CS…)



Our Roadmap's Keystone: Building a reliable Master Pilot for all future projects

The Master Pilot:

A Multi-Variant, Quality/Productivity Test

Master Pilot for MT Evaluation

Build Stage

- Partnering with our internal client (Customer Support) and external vendors (Kantan)

- Building and tuning SMT and NMT systems

Evaluation Stage

Principles:

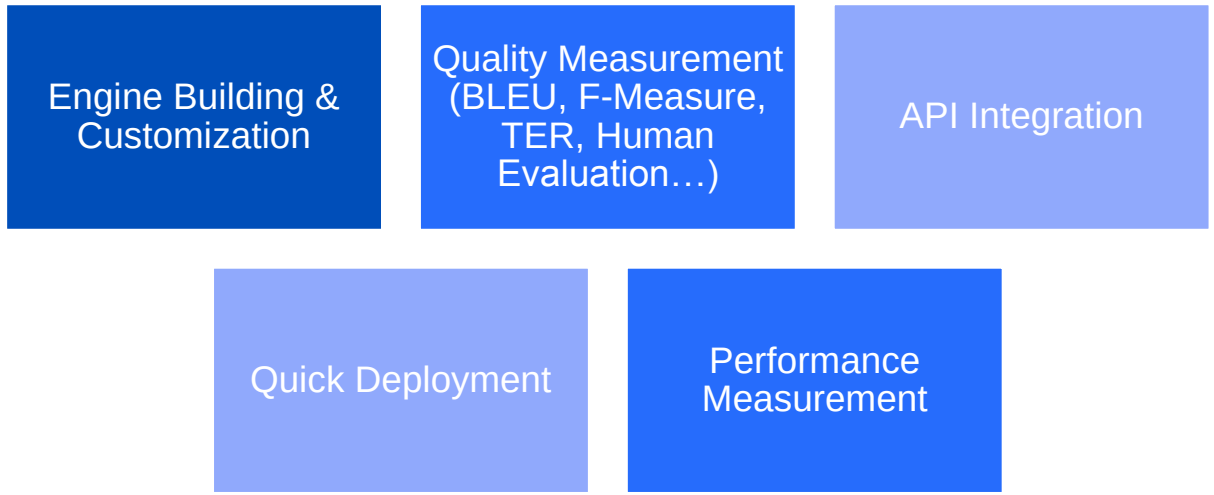
Multi-dimensional:
- Error Analysis
- Quality and Productivity
- Data Correlation

Conclusions

For the pilot: Best engine?
For future pilots: Best process & KPIs?
For the industry:
- Best evaluation method? (Or combination thereof)

For eBay L10n:
How to engage linguists and best leverage their skills?

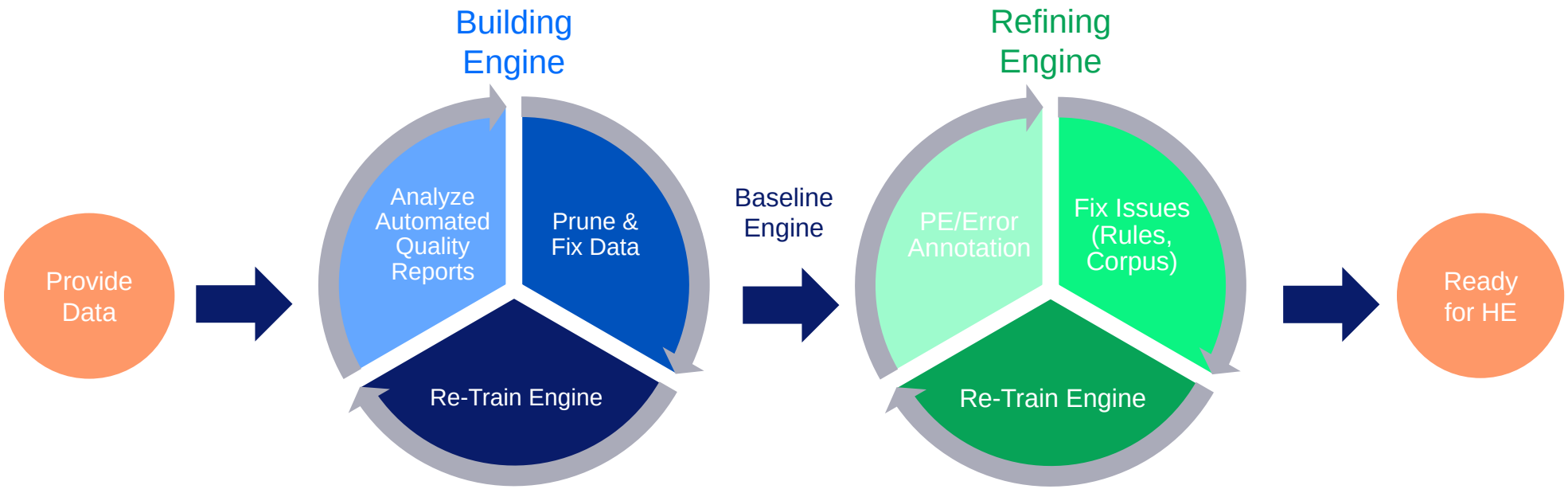
Factors that Decided Us for Our Partner - KantanMT



Phase I:

Engine Building & Report-Based Evaluation with Kantan

Building & Evaluating Engines – The Workflow

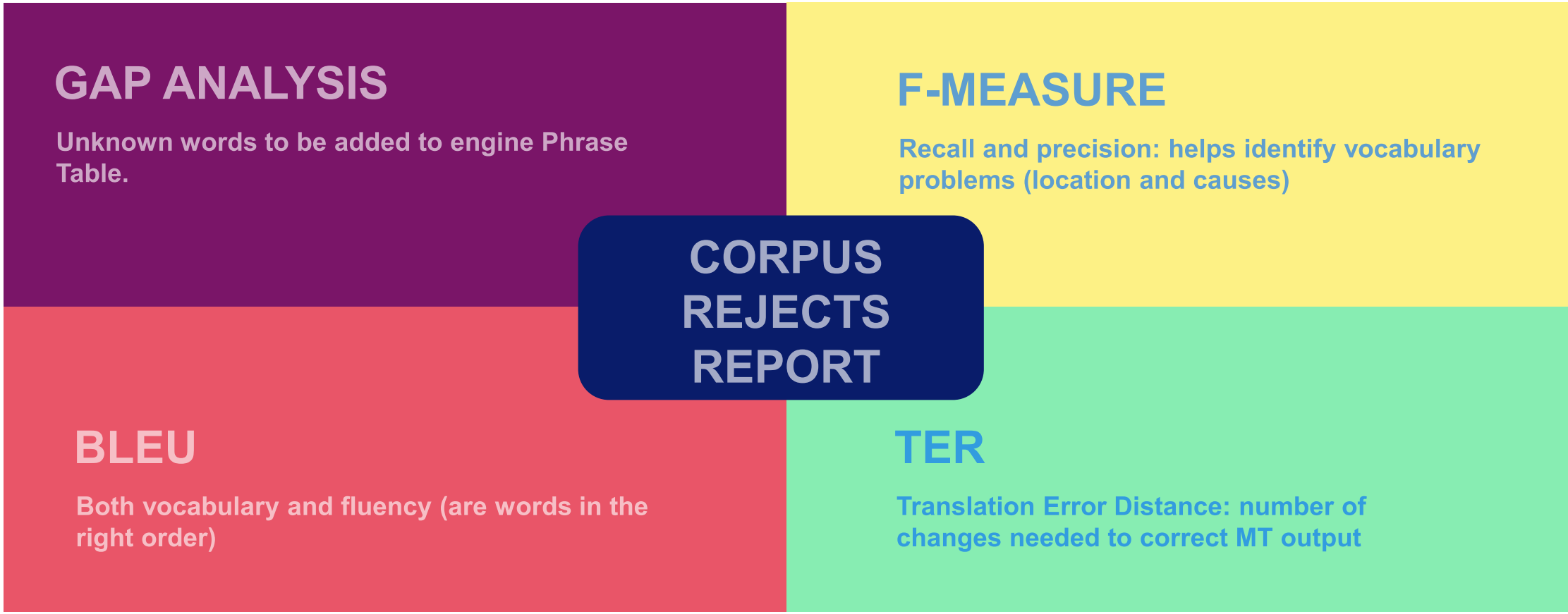


WE FOLLOWED THIS PROCESS FOR BOTH PHRASE-BASED AND NEURAL MT SYSTEMS

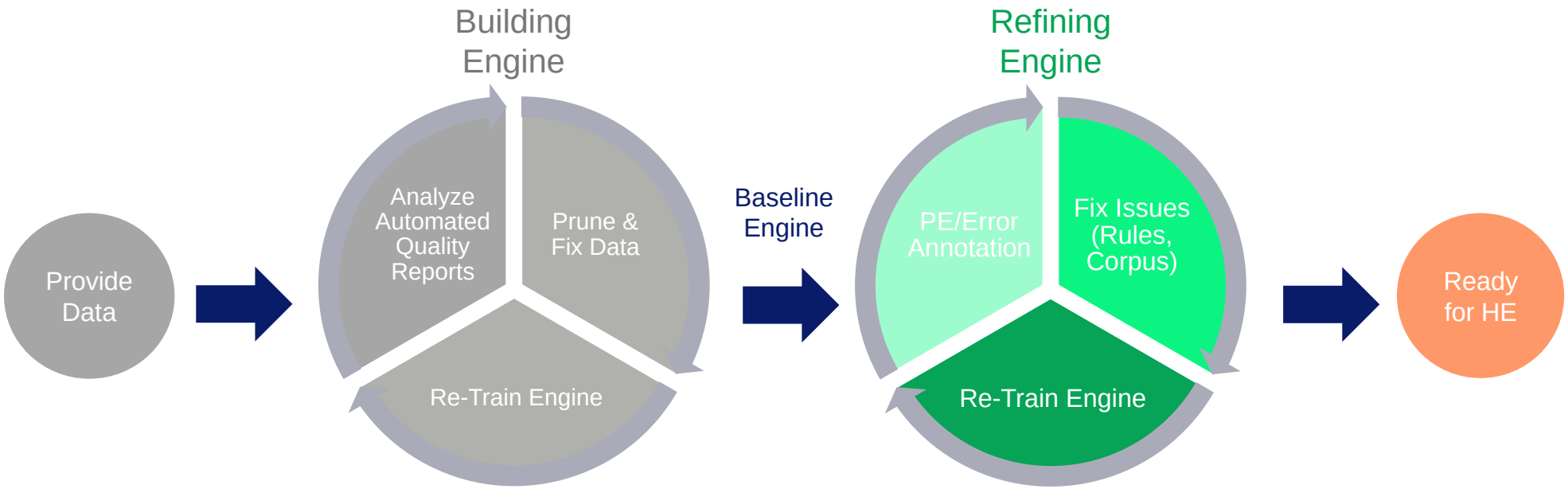
Baseline Engine – Evaluation Based on Automated Reports

Reports produced by:
- Vetting training corpora
- Comparing MT output with the human-translated Reference.

Goal: Finding and fixing major errors to reach threshold scores for Baseline Engine.



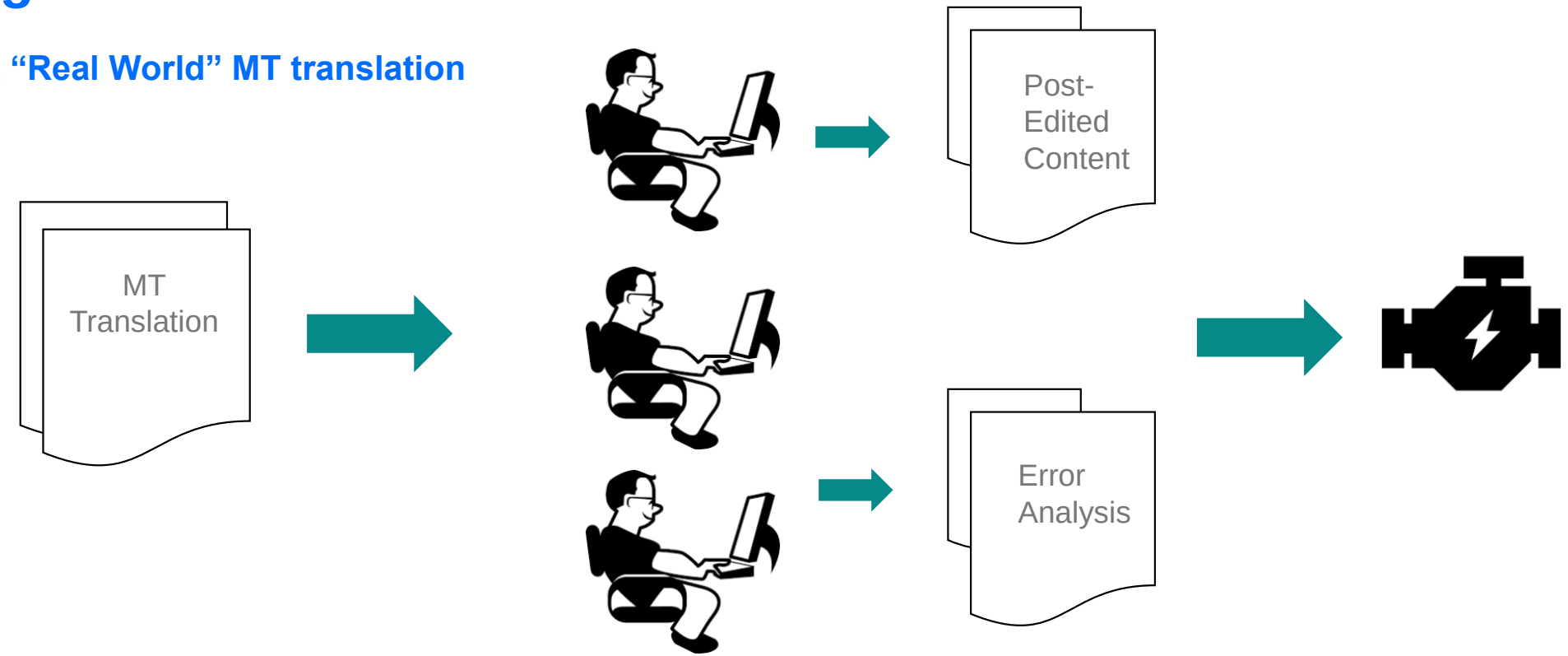
Engine Refinement – Linguistic Quality Review



NOW WE HAVE A BASELINE ENGINE READY, WE HAVE EXPERT LINGUISTS PERFORM A MORE GRANULAR EVALUATION, IN 2 STAGES.

Engine Refinement - Details

First “Real World” MT translation



- 3 EVALUATORS: 2 L10N LINGUISTS AND 1 FINAL CLIENT (CS) REPRESENTATIVE
- 2 ROUNDS TO REACH ACCEPTABLE OUTPUT FOR BENCHMARKING

Engine Refinement – An Effective Error Typology

Error Typology for MT-translated content (DQF-MQM customized subset)

Category	Sub-category	Definition	Action
Terminology		Terminology issues relate to the use of domain- or organization-specific terminology	Add more terms to glossary / add new glossaries
Accuracy	Omission	Translation omits source information	Find out why MT omits information
	Do-not-translate	Term that should stay untranslated is translated	Add terms to NTA list /Tag them in pre-processing
	Untranslated	Term that should be translated stays untranslated	Find out in what areas; we may need additional corpora (what kind?)
	Mistranslation	Term incorrectly translated	Find out whether there is a pattern
Fluency	Grammar - word form	Morphological problem - E.g. “has becomed” instead of “became”.	Fix in corpora / with PEX rules
	Grammar - word order	Bad word order	Fix in engine / with PEX rules
Locale	Format problems - measurement, currency, date/time, address, telephone...	The text does not adhere to locale-specific mechanical conventions and violates requirements for the presentation of content in the target locale.	Fix with PEX rules

ebay

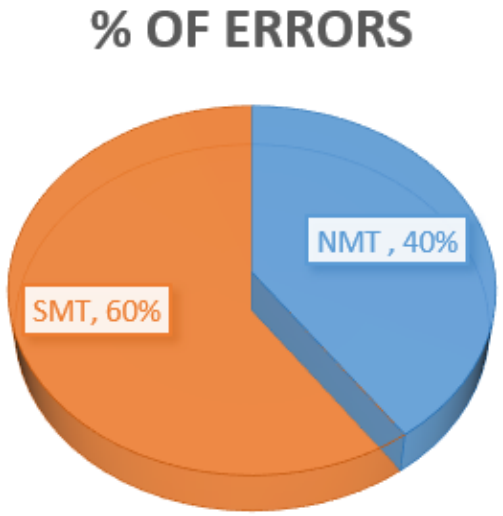
Engine Refinement – An Effective Error Typology

Error Typology for Source Content (DQF-MQM customized subset)

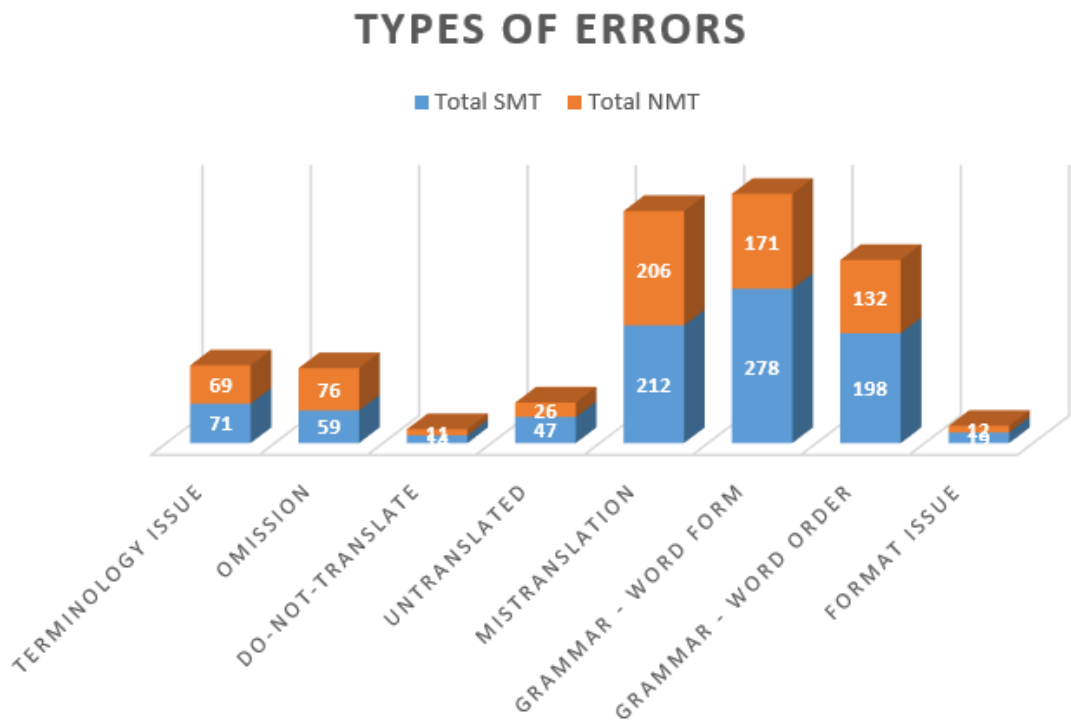
Category	Sub-category	Definition	Action
Ambiguity		The text is ambiguous in its meaning.	Look for a pattern – always identify the error cause when possible. Examples: - Misused punctuation (e.g. “we had problems, coming home” vs “we had problems; coming home”; “high end designer item” vs “high-end designer item”) - Overuse of the -ing form (“I will want you to study after watching TV” can mean “after I watch TV” or “after you watch TV”) - Wrong capitalization (e.g. with a UI element: “Employment Fraud” vs “employment fraud”. Makes it difficult to recognize if this is a UI element (and should stay in English) or not) - Others
Grammar		Function words, word-form, word-order. Typos affecting MT translation.	Look for a pattern (gender/number disagreements, incorrect word order that may cause MT problems) Examples: - high end designer item vs high-end designer item -> Missing hyphen - 3day duration -> Missing space grammar error
Terminology		Inconsistency - multiple words for one concept. Lack of consistency may produce incorrect MT translations, especially in Neural MT.	Provide recommended term.
Design - Markup	**Markup**	Issues related to “markup” (codes used to represent structure or formatting of text, also known as “tags”). Wrong markup can cause tags to be exposed for translation, or missing, which causes a loss of meaning.	Report for content creators to fix. When in doubt as to whether the missing content is a placeholder, use the Ambiguity error type. Examples: - Full URLs: “ATO %20UK%20Communication%20Preferences%20Change.png" />” - Missing placeholders: “Actively selling when occurs”

ebay

Engine Refinement Results – SMT vs NMT Errors



Total errors	NMT	SMT
1501	603	898
	40%	60%



CONCLUSIONS:

NMT produces considerably less errors than SMT

NMT matches or beats SMT in all areas except omissions

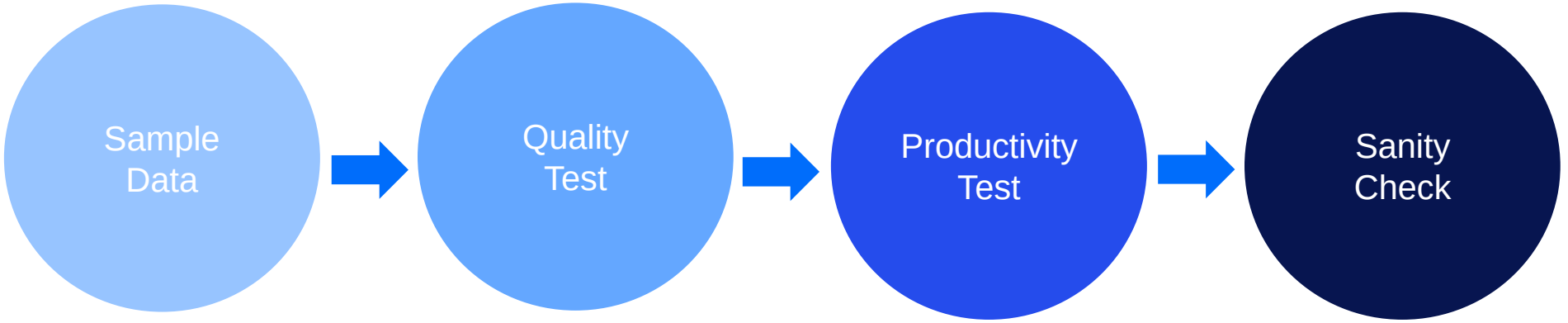
NMT performs specially well in grammar (morphology, word order), i.e. Fluency

ebay

Phase II:

Human Evaluation: Benchmarking SMT vs NMT vs HT

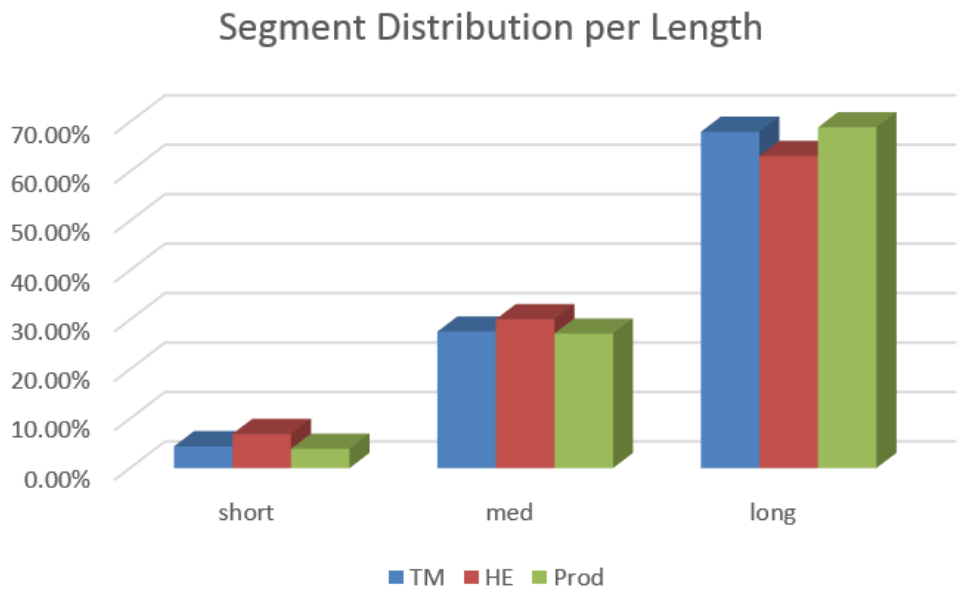
Benchmarking Flow – SMT, NMT and HT



Features	800 representative segments	1-5 Scale Blind randomized test **NMT vs SMT vs HT**	A/B Test (Human Translation vs PE) **Winner MT** vs **HT**	1-5 Scale Linguistic Quality Assurance
Data Points	3 segment lengths (long, medium, short)	Adequacy Fluency Overall Quality	Time spent - HT Time spent - PE PE ED	Final Quality Score

ebay

Data for Quality and Productivity: A Representative Sample



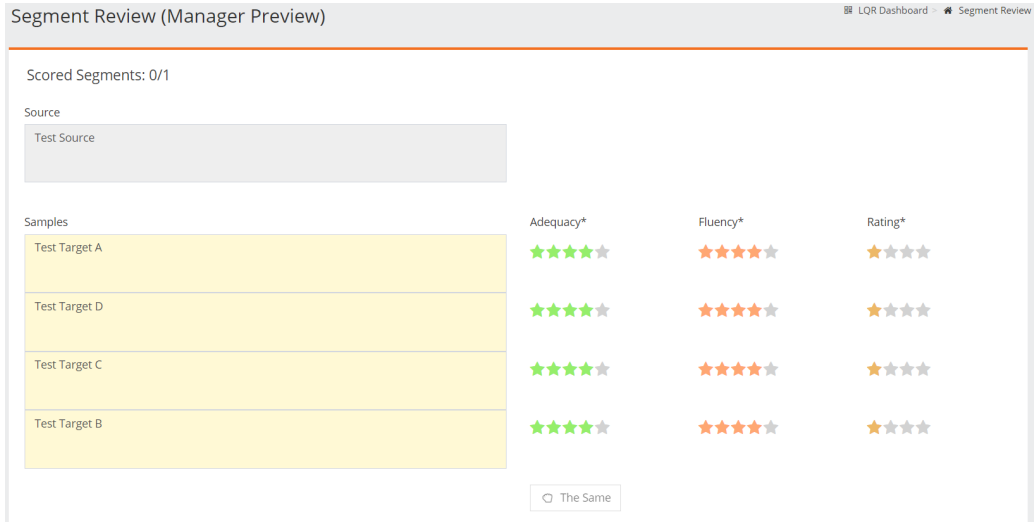
By Silvio Picinini, eBay BPT MTLS

Our sample mirrors the CS TM length distribution:
- Short segments (1-4 words): little context
- Medium segments (6-12 words) simple full sentences
- Long segments (13-35 words) complex sentences

5 sets of short-medium-long segments:
- 2 for post-editing
- 1 for human translation (to compare with PE)
- 1 for human evaluation

Benchmarking: Quality

Quality Evaluation Stage



WHO

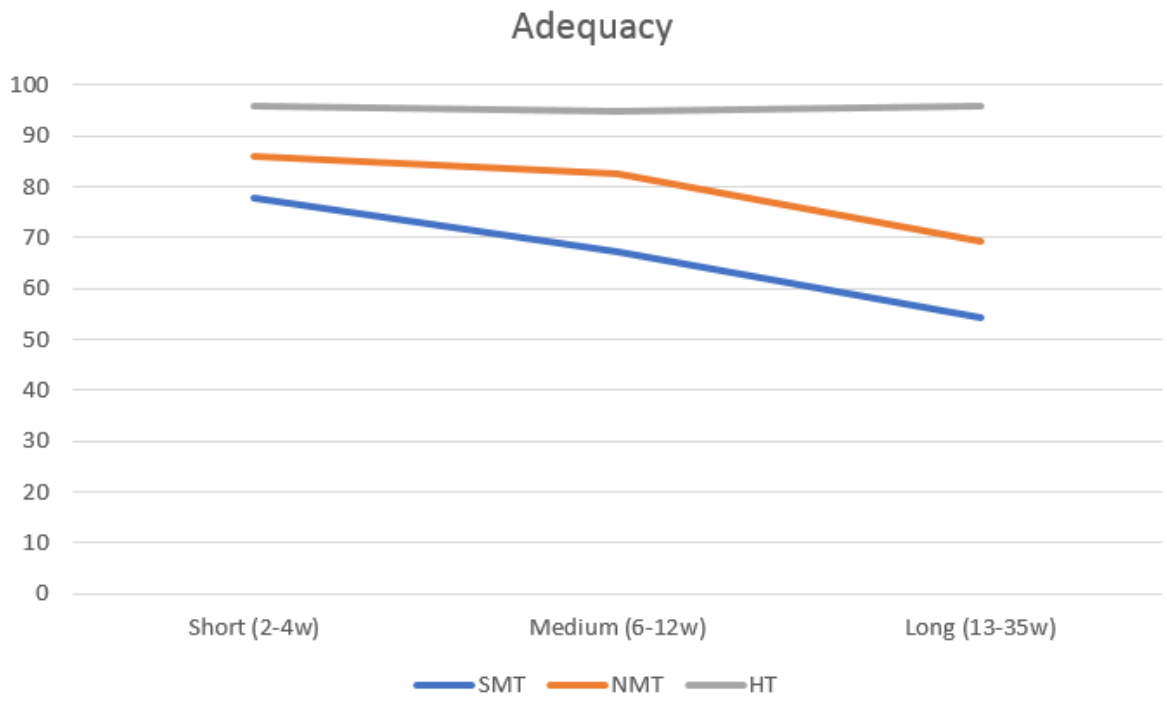
4 Linguists:
- **1 External Vendor**
- **2 eBay In-House Linguists**
- **1 Customer Support**

WHERE

Kantan AB Test Tool:
- Simple, easy-to-use ranking and rating features

WHAT
- NMT vs SMT vs Human Translation
- Adequacy: How much of the source meaning is preserved in the translation
- Fluency: To what extent is the translation grammatical and natural-sounding.
- Overall: General impression

Adequacy Results: Quality per Segment Length

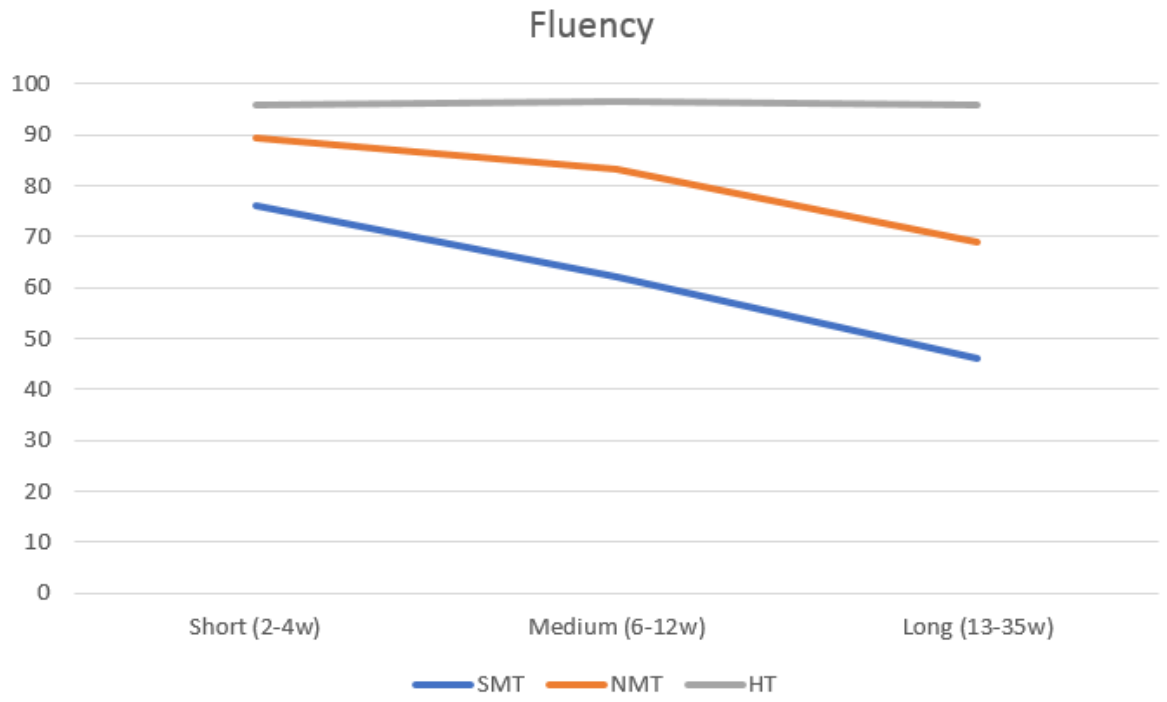


1-100 Scale

- HT Stable high quality (as expected)
- On average, **NMT 22% better than SMT** (79% vs 65%)
- SMT and NMT adequacy declines with longer segments
- NMT is (surprisingly) better **even in shorter segments**

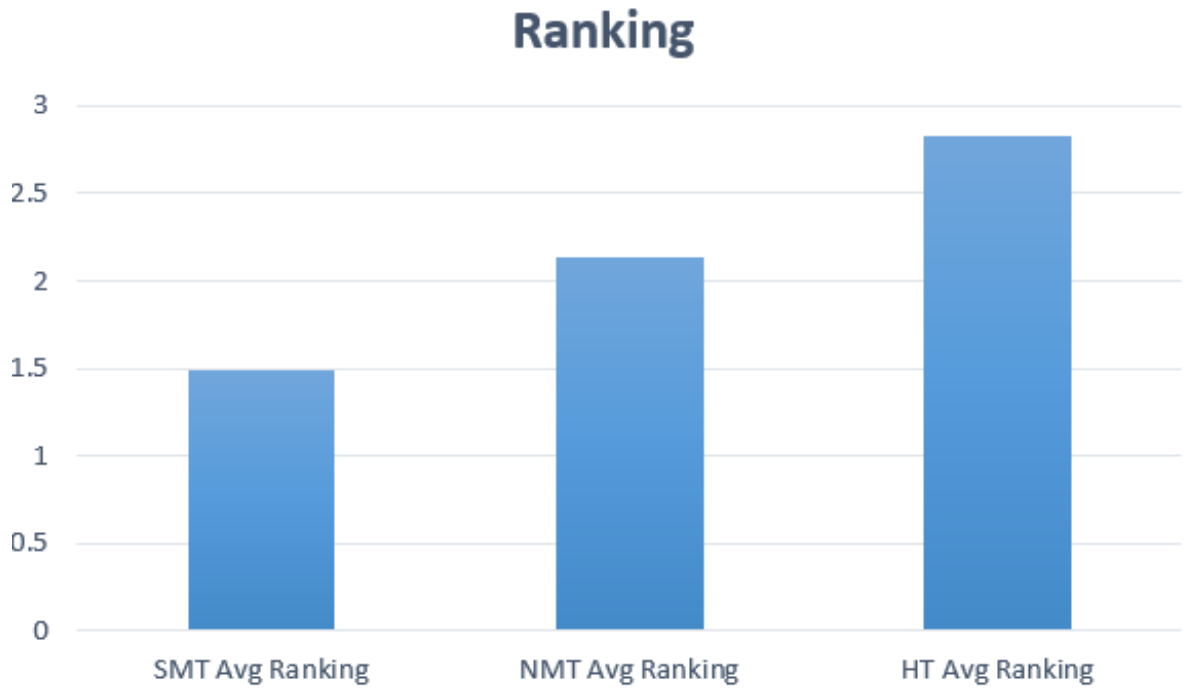
ebay

Fluency Results: Quality per Segment Length



1-100 Scale
HT Stable
On average, **NMT 33% better than SMT** (80% vs 60%)
SMT and NMT adequacy also declines with longer segments **(but NMT holds better - expected)**

Overall HE Ranking



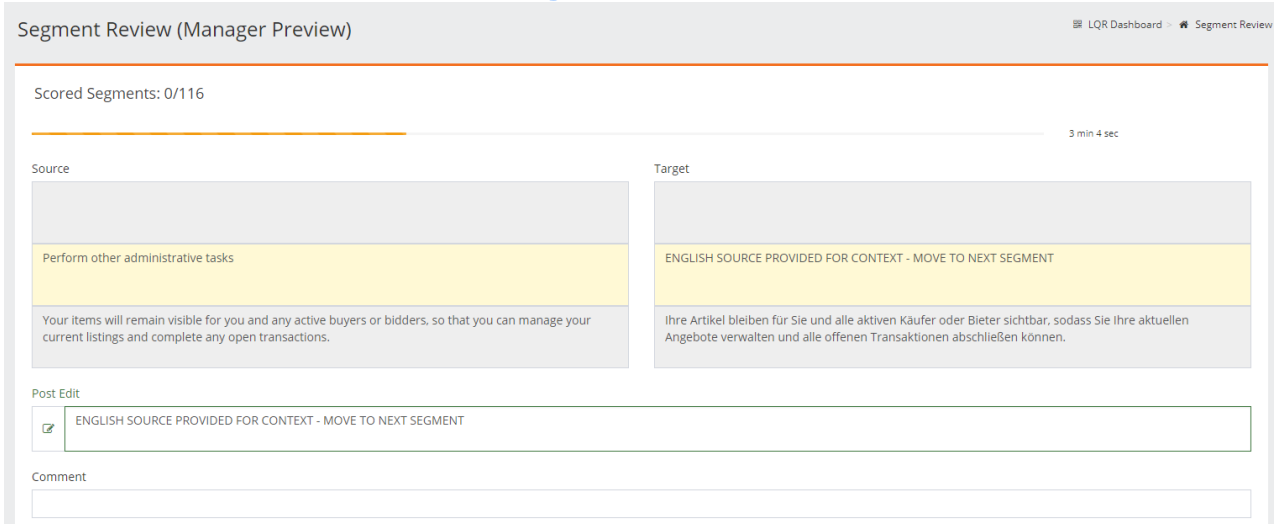
SMT Average Ranking	NMT Average Ranking	HT Average Ranking
1.49 (50%)	2.13 (71%)	2.83 (94%)

By including HT in test set, we determine ideal baseline is 94% of a perfect score

ebay

Benchmarking: Productivity

Productivity Evaluation Stage



WHO

3 Linguists: - 1 External Vendor
- 2 eBay In-House Linguists

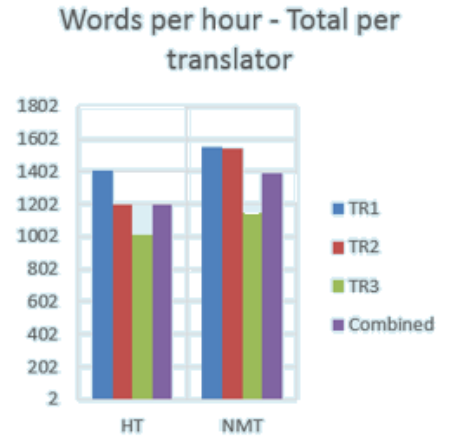
WHERE

Kantan LQR:
- Simple, provides glossary, no TM
- Provides context
- Allows us to track time and edit distance

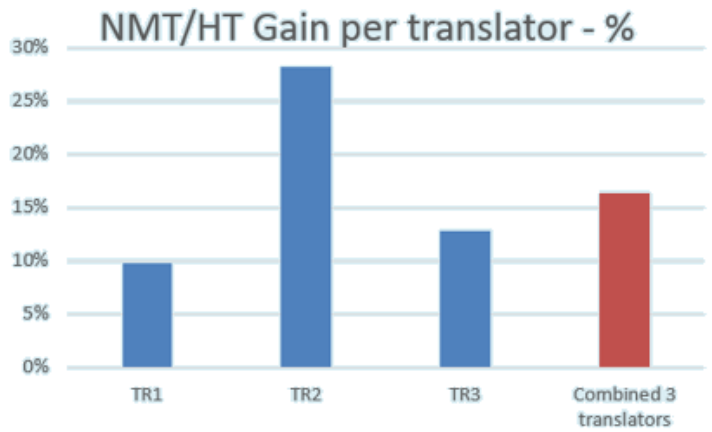
WHAT

- NMT vs Human Translation
- A/B productivity test: linguists translate and post-edit equal parts of a file
- High quality expectation

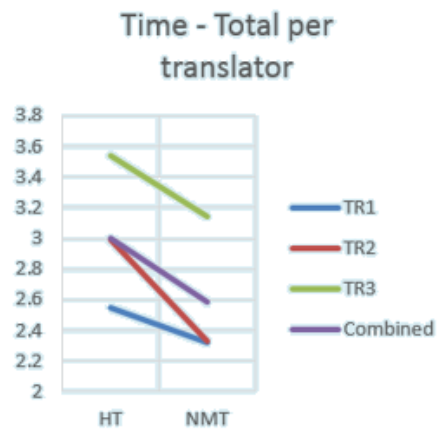
NMT vs HT – Time Gains



NMT Productivity gain: absolute words



NMT Productivity gain: % over HT

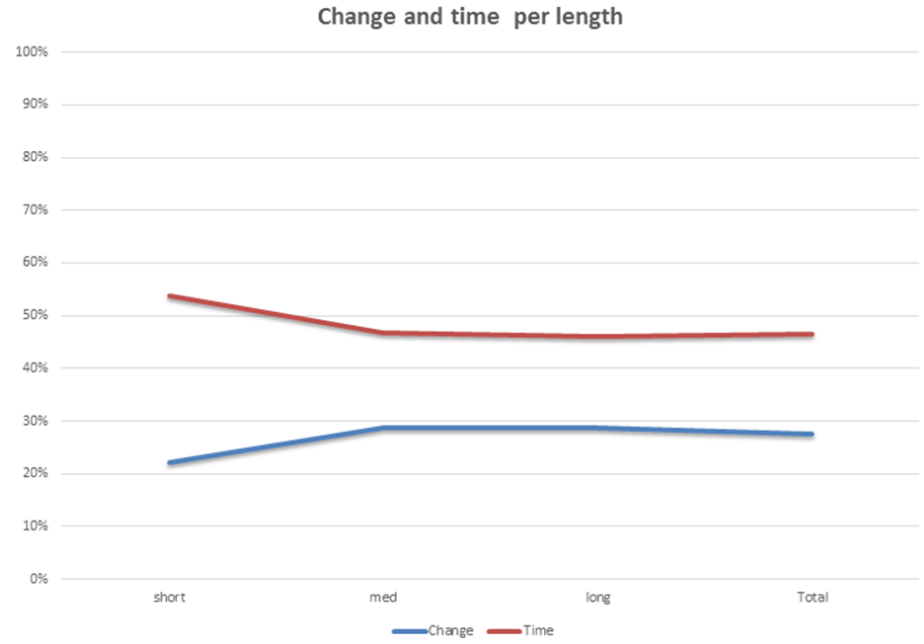


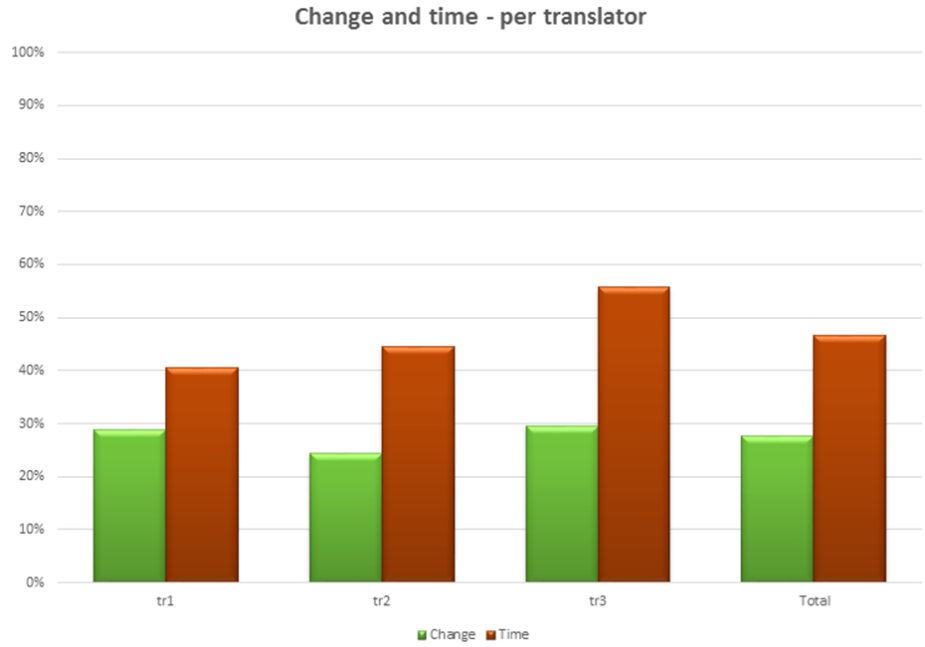
NMT Productivity gain: words/second

PENMT consistently increases productivity (10-27%)

2 in-house translators (1 in particular) leverage greatest gains

NMT vs HT – Correlation Time-Edit Distance





PER SEGMENT LENGHT

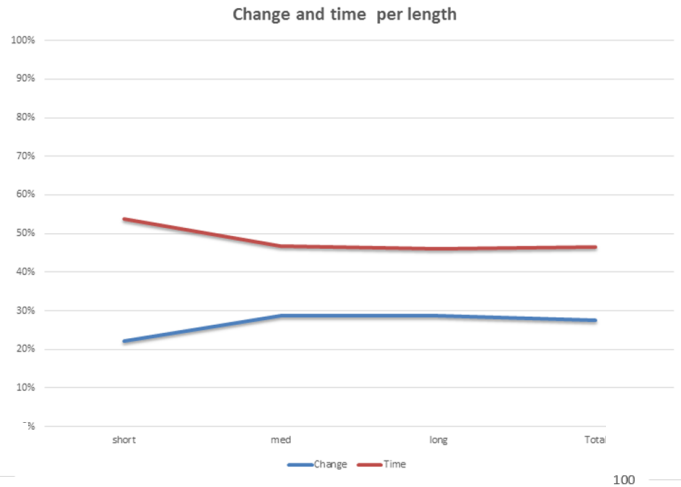
A uniform ratio between edit distance and time to edit, **except** for very short segments, that require proportionally more time (likely significant terms, requiring more research)

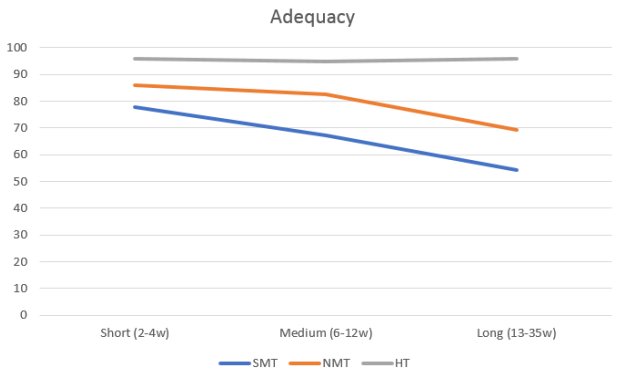
PER TRANSLATOR

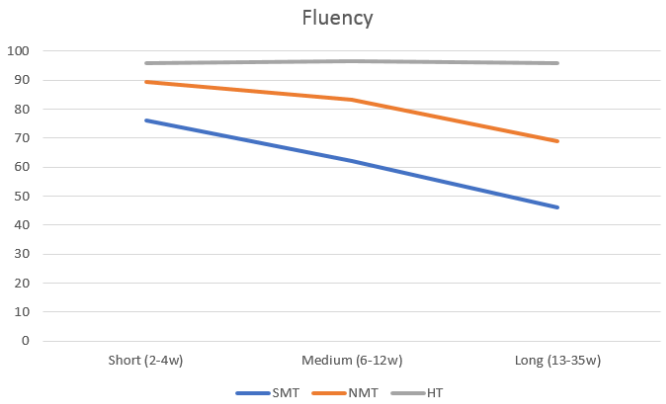
ED and time are mostly aligned, with one exception. one of the linguists's (vendor) time to edit is an outlier.

ebay

NMT vs HT–Correlation Time-Edit Distance vs Adequacy-Fluency

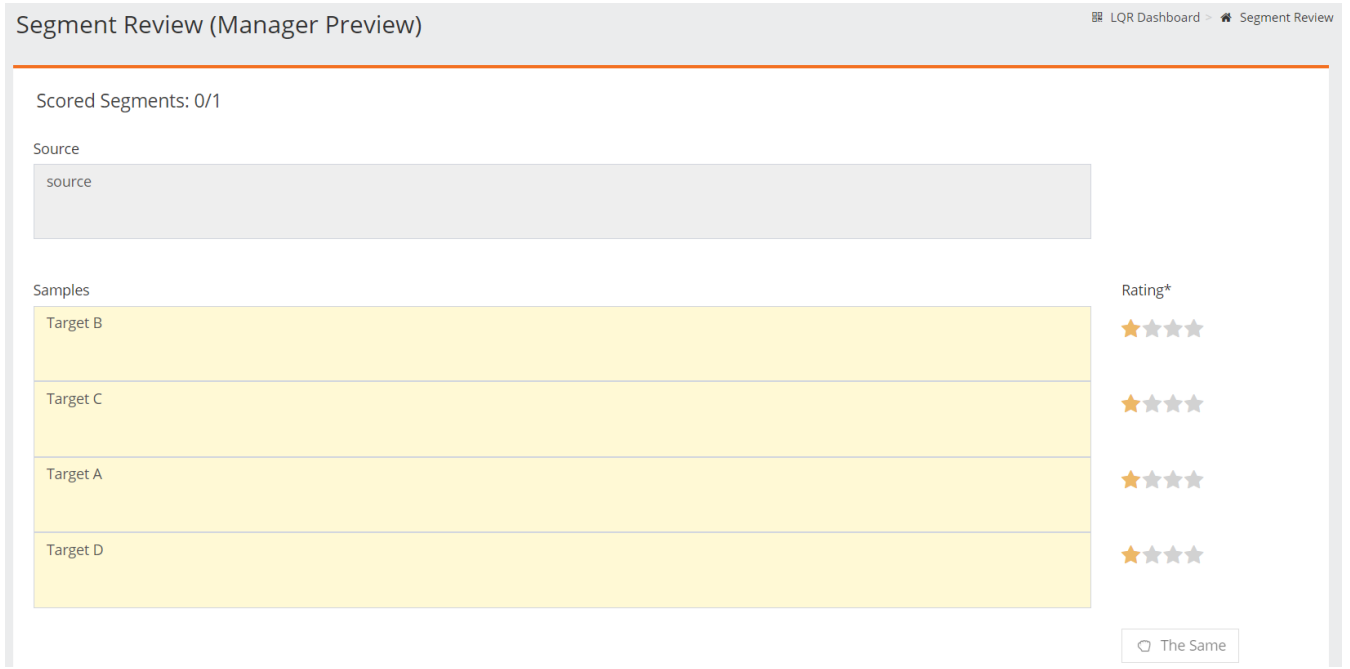






Interestingly, the perceived decline in Adequacy and Fluency for long segments is not reflected in a higher ED or longer time to edit.

Quality Assessment: The Sanity Check



From KantanLQR

A Quality Assessment of post-editors' final quality

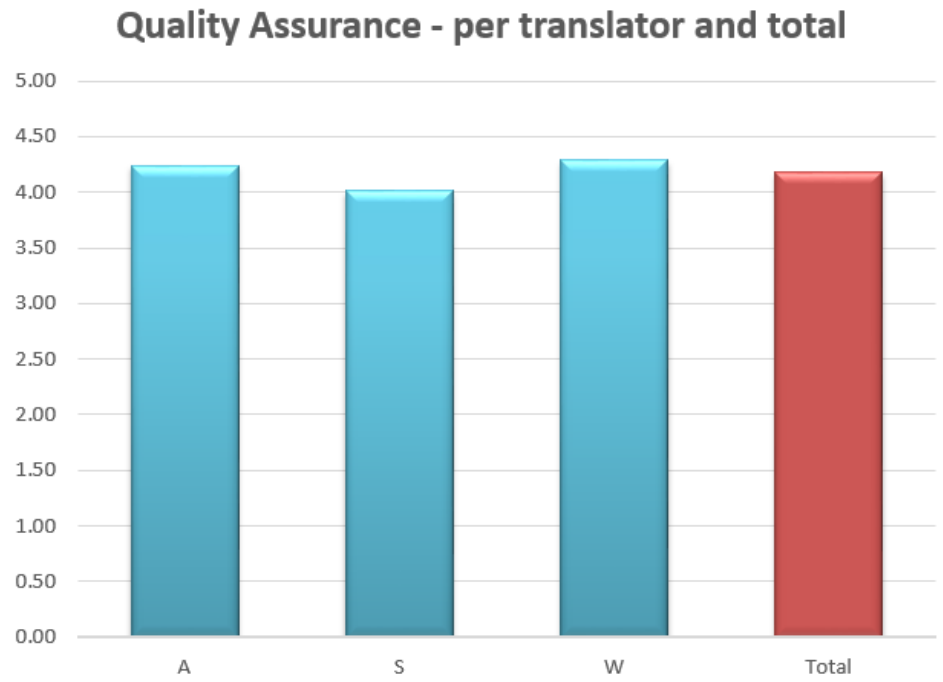
Quality Assessment: Results

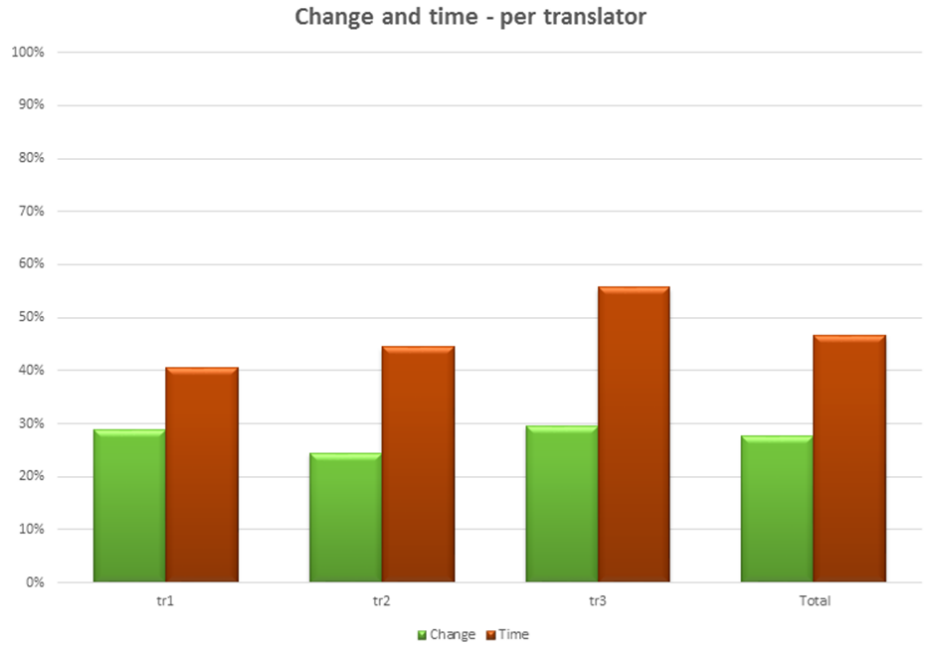
A linguist reviewed a sample of the post-edit work of the evaluators

Quality was very similar: 4.24 - 4.01 - 4.29

Additional Insights

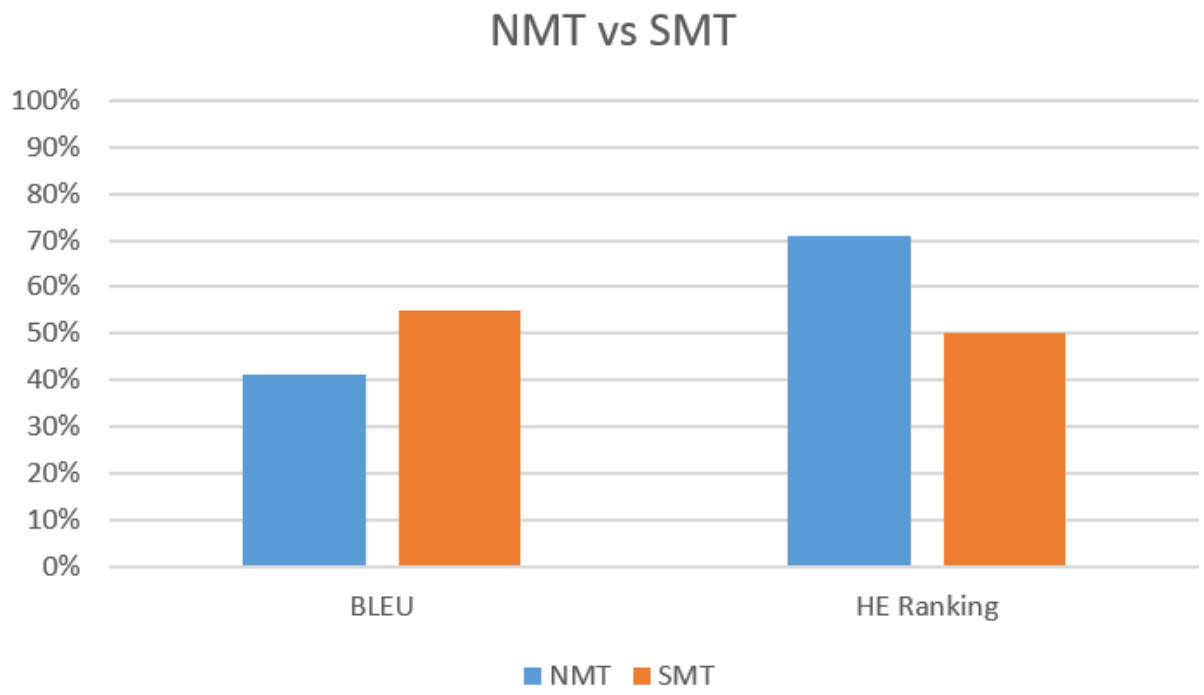
Correlation 1: Outliers in Quality – Edit Distance – Time





Similar quality, similar edit distance, one outlier in time spent:
Further training on post-editing may be useful

Correlation 2: HE shows BLEU bias against NMT



	NMT	SMT
BLEU	41%	55%
HE	71%	50%

Feedback from Participating Linguists

We surveyed all 4 linguists involved in the pilot:

Lessons learned:

- Ensure good communication:
  - Initial presentation with high-level goals
  - For every stage, clear statement of goals and expectations
  - Clearly defined key terms (BLEU, ranking, rating, A/B test…)

- Provide sufficient context for HT/PE (no random strings, enough strings before and after)

- Minimize the number of variables: Use simple tools and basic resources (drop TM, use basic instructions)

ebay

Conclusions

What We Found:

PILOT GOAL

Which is the best engine?

- For the final user: **NMT**
  For the post-editor/vendor: **NMT**

RESEARCH GOALS

- Is BLEU equally reliable for SMT and NMT? **NO**

- Is there a difference between perceived quality and PE effort? **YES**
- Segment length – HE quality:
  Does length affect adequacy/fluency **YES**
  Does NMT and SMT quality vary per segment length **YES**

ORGANIZATIONAL GOALS

- Which are the best roles for each of the stakeholders?
  - MT Vendor: Engine background support
  - eBay MTLS: engine creation, data curation, supporting/training LS for these roles
  - eBay regular LS (for now): quality evaluation

ebay

Questions?