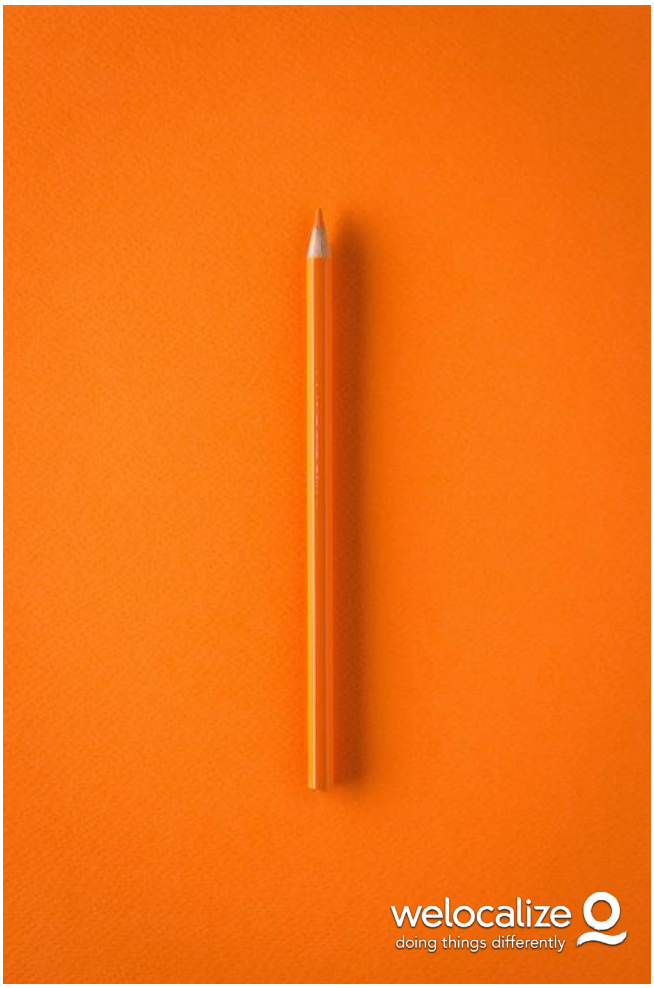
Comparative Evaluation of NMT with Established SMT Programs

Lena Marg
Naoko Miyazaki
Elaine O'Curran
Tanja Schmidt

welocalize
doing things differently

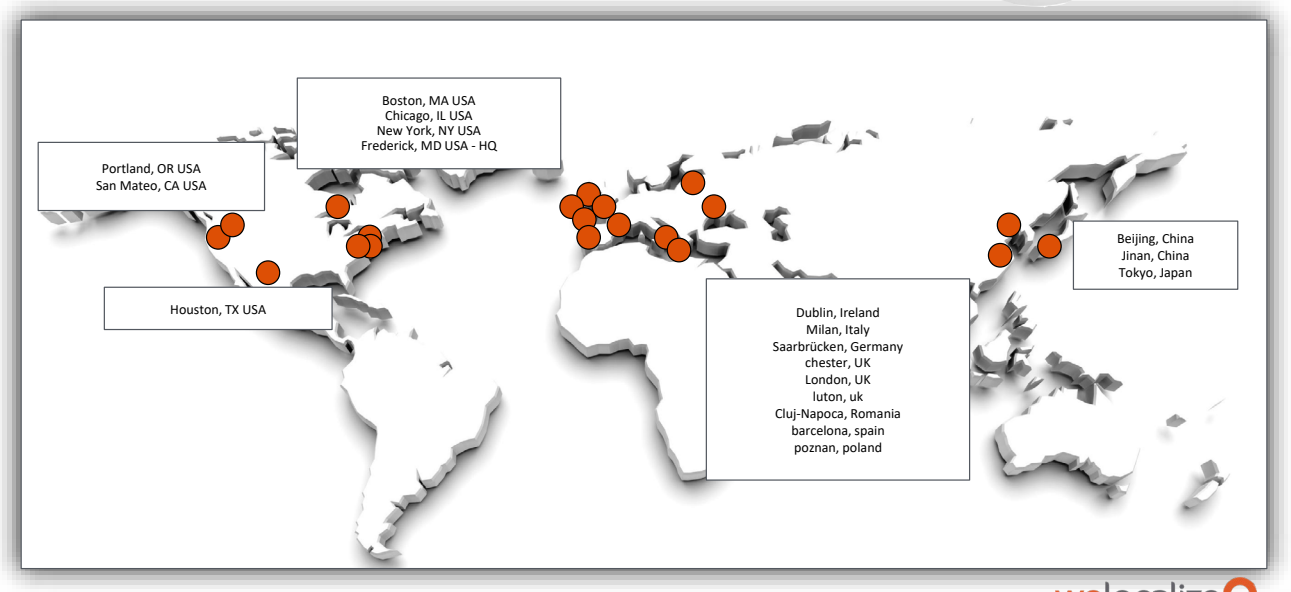
AGENDA

1. Objective
2. Scope of the evaluation
   - Language pairs
   - Content types
   - Size and integrity of the test sets
3. Evaluation methodologies
   - Human evaluations
   - Automatic scoring
4. Results
5. Conclusions



WHO + WHERE WE ARE

WORDS TRANSLATED 2015: 1.15 BILLION
LANGUAGES TRANSLATED: 175+
EMPLOYEES: 1000+
GLOBAL OFFICES: 21
7TH LARGEST PROVIDER IN THE WORLD
4TH LARGEST LSP IN THE US
*2016 Common Sense Advisory



welocalize
doing things differently

Objective

Objective

- Compare the performance of two public NMT systems with a customized SMT solution that is applied in production for two enterprise-level clients.
- Evaluate how generic NMT performs out-of-the-box for different languages and content types that are in high demand in our industry.
- Enable us to make well-founded business decisions as we move forward with our MT strategy.
- Provide data-driven advice and support to our clients.

Scope of the Evaluation

Sampling and Sample Size

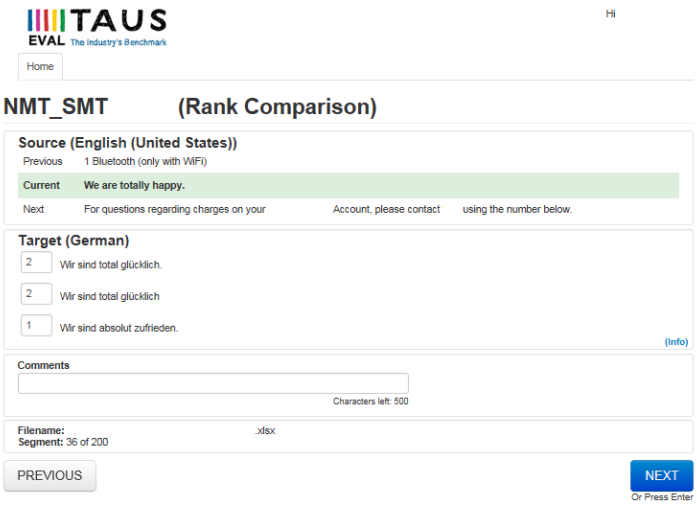
Evaluation Type	Sample Size (TUs)	Sample Origin
Autoscoring (HT)	Approx. 2500	This is the randomized, blind test set taken from the customized SMT engine. The segments in the test set are not included in the engine's training data and originate from production TMs.
Side-by-side engine ranking	200	The 200 segments for human evaluation are randomly selected from the 2500 TU test set described above
Adequacy and Fluency scoring	100	From the 200 segments above, we randomly selected 100 segments for the more detailed human analysis and post-editing sample
Strength and Weaknesses Assessment	100	Same sample as above
Autoscoring (PE)	100	Same sample as above is post-edited and scored

Scope Overview

Evaluation Type	MT Systems	Content Type	Language Pairs	Evaluators
Autoscoring (HT)	Customized SMT, Generic1 NMT, Generic2 NMT	Light Marketing, Technical Documentation	de-DE, fr-FR, ja-JP, pt-BR, ru-RU, zh-CN	Proprietary scoring tool (wescore)
Side-by-side engine ranking	Customized SMT, Generic1 NMT, Generic2 NMT	Light Marketing, Technical Documentation	de-DE, fr-FR, ja-JP, pt-BR, ru-RU, zh-CN	Two evaluators: one account translator, one experienced MT evaluator
Adequacy and Fluency scoring	Customized SMT, Generic2 NMT	Light Marketing, Technical Documentation	de-DE, ja-JP, pt-BR	One evaluator: account translator
Strength and Weaknesses Assessment	Customized SMT, Generic2 NMT	Light Marketing, Technical Documentation	de-DE, ja-JP, pt-BR	One evaluator: account translator
Autoscoring (PE)	Customized SMT, Generic1 NMT	Light Marketing	de-DE, ja-JP, pt-BR	One evaluator: account translator

Side-by-Side Engine Ranking

- The TAUS DQF tool used for this evaluation randomizes the order in which the target segments from the engines being compared are presented. This means the evaluator(s) do not get conditioned into giving anticipated rankings
- Ranking (1,2,3) of the 3 engines, from best to worst
- Allows equal ranking of two or three outputs



Adequacy and Fluency Scoring

Adequacy Score Evaluation Criteria	
5	All meaning expressed in the source appears in the translation. You do not need to refer to the source to understand the meaning.
4	Most of the source meaning is expressed in the translation. You can understand most of the meaning without referring to the source.
3	Much of the source meaning is expressed in the translation. Roughly half the MT output can be understood without referring to the source.
2	Little of the source meaning is expressed in the translation. Although you can guess fractions of the MT output, you cannot understand it without referring to the source.
1	None of the meaning expressed in the source is expressed in the translation. You cannot make any sense of the MT output alone AND/OR the MT output says exactly the opposite of the source.

Fluency Score Evaluation Criteria	
5	Native language fluency. No grammar errors, good word choice and syntactic structure. No PE required.
4	Near native fluency. Few terminology or grammar errors which don't impact the overall understanding of the meaning. Little PE required.
3	Not very fluent. About half of translation contains errors and requires PE.
2	Little fluency. Wrong word choice, poor grammar and syntactic structure. A lot of PE required.
1	No fluency. Absolutely ungrammatical and for the most part doesn't make any sense. Translation has to be re-written from scratch.

Ranking Strengths and Weaknesses

WHICH TRANSLATION IS BETTER WITH REGARD TO:
accuracy (accurate rendition of source meaning)
fluency & style
general domain terminology
client-specific terminology & instructions
completeness (all key information from source is rendered)
redundancy (translation contains additional information not contained in the source)
syntax
grammar
localization (correct format of punctuation; spacing; dates & time, units measurement)
tags & placeholders
spelling
Other

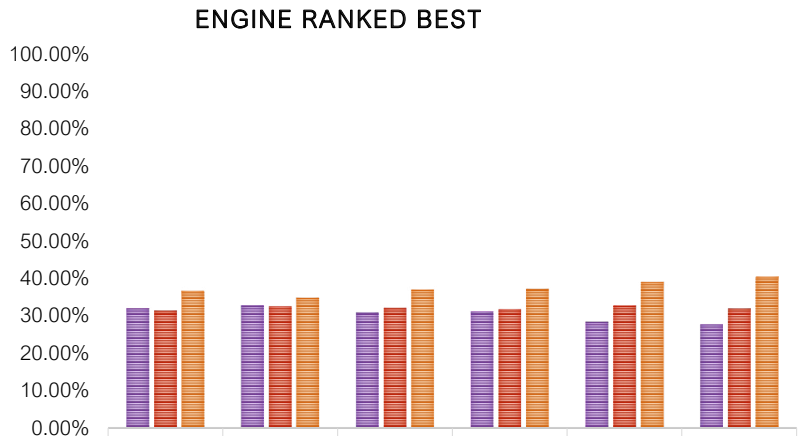
Autoscoring

- BLEU
- NIST
- METEOR
- GTM
- Precision
- Recall
- TER
- PE Distance*

*In our analysis we focus on PE distance, which applies the Levenshtein algorithm and is character-based. Compared to word-based scoring, this method captures morphological post-edits, such as fixing word forms, and we have found it to correlate well with human judgment.

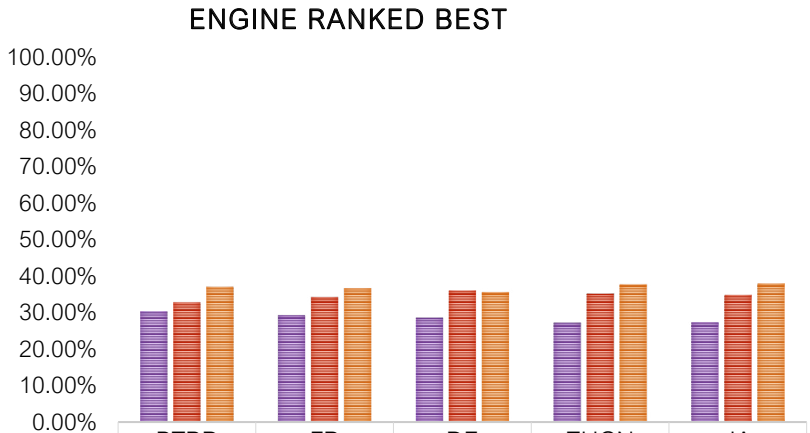
Results

Engine Ranking Results for Light Marketing



	PTBR	FR	DE	RU	ZHCN	JA
Customized SMT	32.02%	32.77%	30.93%	31.18%	28.38%	27.72%
Generic2 NMT	31.37%	32.49%	32.12%	31.68%	32.66%	31.91%
Generic1 NMT	36.61%	34.74%	36.95%	37.14%	38.95%	40.37%

Engine Ranking Results for Technical Documentation



	PTBR	FR	DE	ZHCN	JA
Customized SMT	30.33%	29.30%	28.56%	27.25%	27.35%
Generic2 NMT	32.68%	34.11%	35.94%	35.10%	34.79%
Generic1 NMT	36.99%	36.59%	35.50%	37.65%	37.86%

German Results

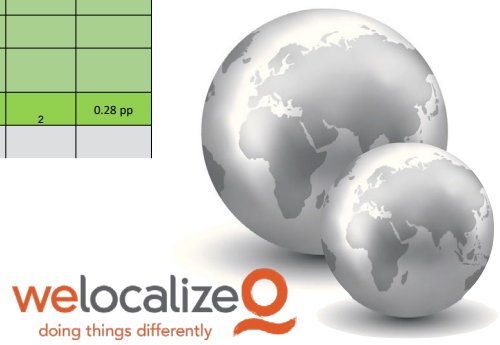
Locale	Evaluation	Light Marketing Content				Technical Documentation Content			
		Generic1 NMT	Generic2 NMT	Customized SMT	Diff Best NMT & SMT	Generic1 NMT	Generic2 NMT	Customized SMT	Diff Best NMT & SMT
de-DE	Ranking	√	2	3	6.02 pp	2	√	3	7.38 pp
	Adequacy			√	0.06		√		0.08
	Fluency		√		0.07		√		0.45
	Accuracy						√		
	Fluency & Style		√				√		
	Syntax		√				√		
	Grammar		√				√		
	Terminology			√					
	Completeness			√			√		
	Localization			√					
	Edit Distance (HT)	2	3	√	3.32 pp	√	3	2	1.12 pp
	Edit Distance (PE)	2		√	1.55 pp				

Japanese Results

Locale	Evaluation	Light Marketing Content				Technical Documentation Content			
		Generic1 NMT	Generic2 NMT	Customized SMT	Diff Best NMT & SMT	Generic1 NMT	Generic2 NMT	Customized SMT	Diff Best NMT & SMT
ja-JP	Ranking	√	2	3	12.96 pp	√	2	3	10.51 pp
	Adequacy		√		0.32		√		0.76
	Fluency		√		0.2		√		0.49
	Accuracy		√				√		
	Fluency & Style		√				√		
	Completeness		√				√		
	Syntax		√				√		
	Grammar		√				√		
	Terminology			√				√	
	Spelling							√	
	Edit Distance (HT)	√	3	2	8.17 pp	√	3	2	5.79 pp
	Edit Distance (PE)	√		2	21.07 pp				

Brazilian Results

Locale	Evaluation	Light Marketing Content				Technical Documentation Content			
		Generic1 NMT	Generic2 NMT	Customized SMT	Diff Best NMT & SMT	Generic1 NMT	Generic2 NMT	Customized SMT	Diff Best NMT & SMT
pt-BR	Ranking	√	3	2	4.59 pp	√	2	3	6.65 pp
	Accuracy		√		0.09		√		0.26
	Fluency		√		0.45		√		0.28
	Accuracy						√		
	Fluency & Style						√		
	Completeness		√				√		
	Redundancy		√						
	Syntax		√				√		
	Grammar		√				√		
	Terminology			√					
	Localization			√					
	Tags & Placeholders			√					
	Edit Distance (HT)	2	3	√	1.68 pp	√	3	2	0.28 pp
	Edit Distance (PE)	2		√	3.62 pp				



French, Russian, Simplified Chinese Results

Locale	Evaluation	Light Marketing Content				Technical Documentation Content			
		Generic1 NMT	Generic2 NMT	Customized SMT	Diff Best NMT & SMT	Generic1 NMT	Generic2 NMT	Customized SMT	Diff Best NMT & SMT
fr-FR	Ranking	√	3	2	1.97 pp	√	2	3	7.29 pp
	Edit Distance (HT)	2	3	√	2.02 pp	2	3	√	0.62 pp
zh-CN	Ranking	√	2	3	10.57 pp	√	2	3	10.40 pp
	Edit Distance (HT)	√	3	2	5.87 pp	√	3	2	3.12 pp
ru-RU	Ranking	√	2	3	5.95 pp				
	Edit Distance (HT)	2	3	√	1.58 pp				

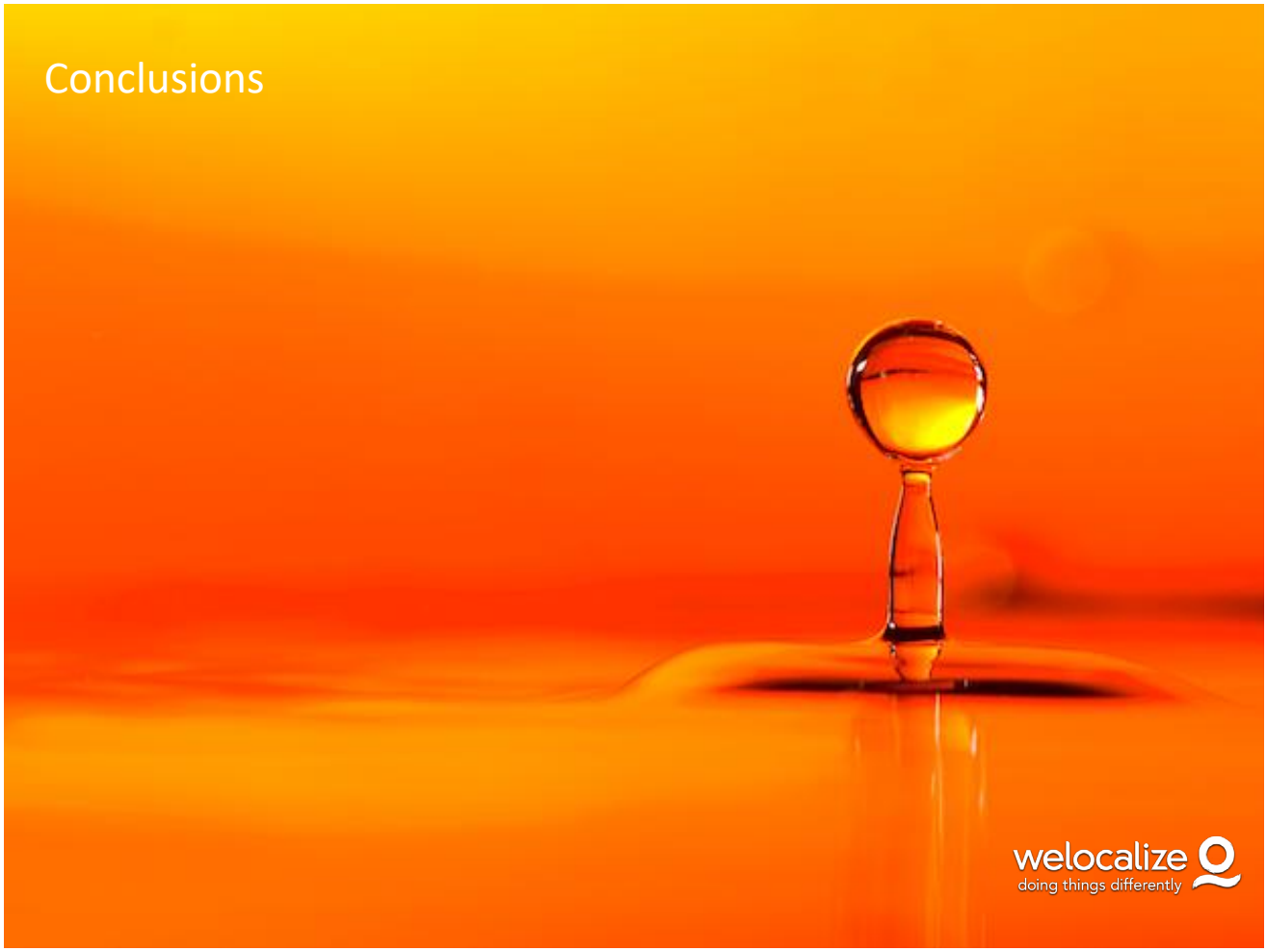
SUMMARY

- All evaluators prefer generic NMT during side-by-side ranking, the first evaluation task.
- NMT also wins Adequacy & Fluency scoring with the exception of German Adequacy for Light Marketing.
- Evaluators for JA, DE, PTBR overall prefer customized SMT for terminology and localization-related issues, but NMT for fluency, style, grammar and syntax. JA also prefers NMT for accuracy.
- NMT outperforms SMT more consistently on Technical Documentation than on Light Marketing.
- For Technical Documentation the autoscores favor NMT, while they show mixed results for Light Marketing.
- After completing the post-editing task on Light Marketing, the German and Brazilian translators had a slight preference for SMT, contradicting the previous human evaluation results and indicating that the autoscores may be more accurate.

SUMMARY

- The most significant quality improvement with NMT are for Chinese and Japanese
- For the other languages, the quality differences between NMT and SMT are less pronounced

Locale	Evaluation	Light Marketing				Technical Documentation			
		Generic NMT1	Generic NMT2	Customized SMT	Diff Best NMT & SMT	Generic NMT1	Generic NMT2	Customized SMT	Diff Best NMT & SMT
de-DE	Ranking	√	2	3	6.02 pp	2	√	3	7.38 pp
	Accuracy			√	0.06		√		0.08
	Fluency		√		0.07		√		0.45
	Edit Distance	2	3	√	3.32 pp	√	3	2	1.12 pp
	Edit Distance (PE)	2		√	1.55 pp				
fr-FR	Ranking	√	3	2	1.97 pp	√	2	3	7.29 pp
	Edit Distance	2	3	√	2.02 pp	2	3	√	0.62 pp
ja-JP	Ranking	√	2	3	12.96 pp	√	2	3	10.51 pp
	Accuracy		√		0.32		√		0.76
	Fluency		√		0.2		√		0.49
	Edit Distance	√	3	2	8.17 pp	√	3	2	5.79 pp
	Edit Distance (PE)	√		2	21.07 pp				
pt-BR	Ranking	√	3	2	4.59 pp	√	2	3	6.65 pp
	Accuracy		√		0.09		√		0.26
	Fluency		√		0.45		√		0.28
	Edit Distance	2	3	√	1.68 pp	√	3	2	0.28 pp
	Edit Distance (PE)	2		√	3.62 pp				
zh-CN	Ranking	√	2	3	10.57 pp	√	2	3	10.40 pp
	Edit Distance	√	3	2	5.87 pp	√	3	2	3.12 pp
ru-RU	Ranking	√	2	3	5.95 pp				
	Edit Distance	2	3	√	1.58 pp				



Conclusions

- Generic NMT is a suitable alternative for generic domains across all the language pairs.
- In the technology domain, generic NMT is a suitable alternative for some language pairs, such as Chinese and Japanese, where we see a substantial increase in performance compared to customized SMT.
- Because most of our enterprise-level programs rely on accurate terminology, we recommend waiting for customized NMT for the remaining language pairs.
- Post-edit distance on actual post-edited content proved to be the most reliable metric in our evaluation. Ranking and Adequacy & Fluency scoring from the same resource was not always consistent. Autoscores (HT) did not correlate with human evaluations in several cases.

NEXT STEPS

We are running several follow-up pilots:

1) Comparing the performance of customized NMT against customized SMT.
2) Comparing Post-edit distance in live production using customized SMT and generic NMT. We would like to see if more extensive production data will confirm our initial findings.