Insights from Gathering MT Productivity Metrics at Scale

Georg Kirchner
Dell Technologies
42 South Street
Hopkinton, MA, 01748, U.S.A.
Georg.Kirchner@dell.com

Abstract

In this paper, we describe Dell EMC's framework to automatically collect MT-related productivity metrics from a large translation supply chain over an extended period of time, the characteristics and volume of the gathered data, and the insights from analyzing the data to guide our MT strategy.

Aligning tools, processes and people required decisions, concessions and contributions from Dell management, technology providers, tool implementors, LSPs and linguists to harvest data at scale over 2+ years while Dell EMC migrated from customized SMT to generic NMT and then customized NMT systems.

For content in two quality tiers, we ranked language pairs by productivity, graphed trendlines, compared the time needed to edit machine translations versus fuzzy matches, studied the time spent on segments with no post-edits, and going by the post-edit density, reviewed segment distribution on a post-edit scale of 1 to 10 and any correlation between the extent of edits and segment length.

1 Gathering Data at Scale

Dell's translation efforts produce significant amounts of linguistic data. Getting to the data, however, is not trivial since it originates with hundreds of linguists who are one or two organizational layers removed in Dell's external supply chain. Each linguist may prefer a different CAT tool, with or w/o the necessary features to track metrics for productivity or quality. Especially if desktop CAT tools require manual configuration from linguists, the constant churn in the resource pool makes it difficult to collect data reliably over time.

For various operational needs, when looking for alternatives six years ago, we qualified translation technology and implemented it as a collaborative environment for linguists to share data in real time. After we had integrated this environment, GlobalLink[1], with the TAUS DQF Dashboard[2], we were ready to harvest metrics on productivity from our Microsoft Translator[3] MT systems automatically between August 2017 and February 2020.

2 The Metrics

We will discuss the following MT metrics: productivity and post-edits. We chose to measure both metrics at the first linguistic step, Translation, although our linguists may have made further changes downstream at the Editing, Proofing, Client Review and Feedback Implementation steps.

The TAUS DQF Dashboard expresses productivity in words post-edited per hour. This number is calculated from the number of words in segments and the milliseconds these segments are active in the CAT tool for editing.

As for post-edits (PED), the TAUS DQF Dashboard distinguishes between post-edit density (PEDe) and post-edit distance (PEDi). Both are calculated with the Levenshtein algorithm (1966).

[1] https://www.translations.com/globallink/index.html
[2] https://qd.taus.net/
[3] https://www.microsoft.com/en-us/translator/business/translator-api/

PEDe expresses changes across the entire sample in percentages, in the average number of characters changed per 100 characters.

At the segment level, PEDi expresses the changes in absolute numbers, i.e., in characters changed per segment; we will call it aPEDi. Normalized to the length of the segment, the PEDi expresses changes on a scale of 0 to 10; we will call it nPEDi. As an example: 10 characters changed in a 20 character-long segment will result in an aPEDi of 10 and an nPEDi of 5.

3 Caveats

As we analyzed the accumulated data for this paper, we found that post-edits made in a single CAT tool session at the translation step correctly capture the full extent of post edits, even if the linguist revisits a segment multiple times in the same session. If the linguist edited the same segment in separate CAT tool sessions, only the edits made in the last session are captured. Because of this, we are underreporting the post-edit distance for an unknown number of segments.

Another caveat is that we decided not to track productivity for human translations (HT). We originally expected that for a given quality tier, we would either MTPE or HT all jobs. And comparing productivity between MTPE and HT jobs across quality tiers would not result in a fair comparison. Later on, we found that our PMs did apply HT workflows selectively to MTPE quality tiers. Had we adjusted our setup, we would now have data to benchmark MTPE against HT productivity.

4 Ranking by Productivity

The most basic exercise is to rank our major language pairs by productivity. These hourly word numbers below result from dividing cumulative MT words by cumulative post-editing time between August 2017 and December 2019.

Good enough		
Souce	Target	Words / Hour
EN-US	pt-BR	2147
EN-US	it-IT	1801
EN-US	fr-FR	1729
EN-US	zh-CN	1409
EN-US	es-MX	1373
EN-US	de-DE	1314
EN-US	ko-KR	1304
EN-US	ja-JP	991
EN-US	ru-RU	905

High quality		
Source	Target	Words / Hour
EN-US	pt-BR	1486
EN-US	fr-FR	1445
EN-US	zh-CN	1204
EN-US	it-IT	1195
EN-US	es-MX	1120
EN-US	de-DE	1040
EN-US	ko-KR	952
EN-US	ja-JP	787
EN-US	ru-RU	743

Table 1: Average number of MT words post-edited per hour and quality tier

The hourly throughput does not account for elapsed time, such as research while segments in the CAT tool are inactive and the time tracker is not running. Therefore these productivity numbers are somewhat theoretical and do not mean that our PT-BR linguists post-edit 2147 x 8 = 17,176 words per day. But they surely suggest that the historic translation output of 2000 words per day is outdated and the actual productivity is significantly higher due to translation technology.

While the language ranking is roughly as expected, there are surprises: ZH-CN ranks relatively high compared to JA-JP and KO-KR; and ES-MX (Latin American Spanish) ranks low compared to FR-FR, IT-IT and PT-BR.

Productivity clearly varies by quality tiers, being higher for Good enough than High quality content. This may be due to varying levels of linguistic complexity and expectations, or the simple fact that Good enough jobs are bigger than High quality jobs, think of product documentation vs. marketing material. And the more volume in a given job, the easier it is for linguists to pick up speed.

EN-US > DE-DE	Good enough	High quality
Total words	1,115,804	2,633,399
# of jobs	484	3,463
Average words / job	2,305	760

Table 2: Average number of new words per job.

Let's see if there is a correlation between BLEU scores generated automatically by Microsoft's MT system customization environment, Custom Translator[4], and our language ranking.

[4] https://docs.microsoft.com/en-us/azure/cognitive-services/translator/custom-translator/overview

Source	Target	Training	Dictionary	BLEU - Dell	BLEU - Baseline
EN-US	ES-MX	1,499,349	0	71.19	49.03
EN-US	PT-BR	1,215,351	2,308	67.16	51.79
EN-US	JA-JP	2,825,902	0	63.02	43.27
EN-US	FR-FR	1,379,887	0	62.58	47.45
EN-US	ZH-CN	1,513,565	2,308	59.93	44.48
EN-US	DE-DE	1,363,042	2,307	58.41	41.36
EN-US	IT-IT	623,477	2,364	56.18	37.68
EN-US	KO-KR	1,099,853	2,364	50.18	32.52
EN-US	RU-RU	556,592	2,364	34.48	20.21

Table 3: Automatically generated BLEU scores during NMT system customization

Training means bi-lingual TMX files containing human translations or post-edited machine translations. The training data is counted in Translation Units (TUs); assume 14 words per TU. *Dictionary* means a phrase table of mostly do-not-translate items such as product names. *BLEU – Dell* is the score based on the customization effort; *BLEU – Baseline* is the Microsoft Translator stock NMT system.

Within the overall correlation between BLEU and productivity, the top-ranking BLEU score for ES-MX failed to predict average post-editing productivity; similarly, the better-than-expected BLEU score for JA-JP failed to predict low post-editing productivity. It has been observed before that productivity and BLEU scores do not correlate necessarily (Koponen, 2016).

There is a caveat to our correlation of productivity and BLEU since productivity was calculated using data collected since August 2017, while the BLEU score is for customized NMT systems deployed only since March 2019.

Productivity ranking of MT systems provides helpful context when triaging linguist feedback on MT output quality, especially when combining the ranking with nPEDi distributions for a particular language or across languages for a given job. Please zoom in.

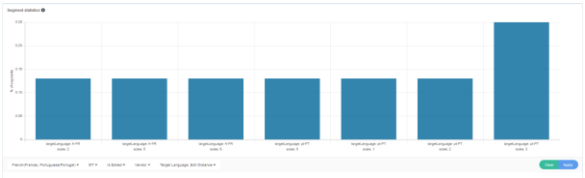

Table 4: Job-level comparison of nPEDi distribution between FR-FR and PT-BR in the TAUS DQF Dashboard.

Also, knowing your numbers allows you to place your MT technology relative to potential alternatives in the market.

5 Productivity and PEDe Trendlines

We wanted to understand if our MT output is getting better, worse or is stable over time.

For this, we compiled productivity and PEDe trends by transcribing monthly averages from the TAUS DQF Dashboard into MSFT Excel and applying linear trendlines. We then noted start and end values of these trendlines along with the number of words machine translated for statistical context.

Target	GE - Words	GE - Prod	GE - PEDe	HQ - Words	HQ - Prod	HQ - PEDe
ZH-CN	1,431,924	20%	-24%	3,232,031	9%	28%
JA-JP	1,607,365	0%	-46%	2,847,134	21%	20%
KO-KR	794,571	-7%	-48%	2,110,117	26%	-15%
DE-DE	1,027,343	33%	-48%	1,862,915	14%	-23%
FR-FR	1,116,151	-5%	-43%	1,814,899	25%	-19%
PT-BR	1,013,459	10%	-29%	1,717,225	14%	-29%
ES-MX	1,147,095	0%	-46%	1,689,249	32%	-28%
IT-IT	549,231	145%	-28%	1,346,025	36%	-22%
RU-RU	530,633	16%	-24%	1,104,929	67%	-6%
NL-NL				520,293	64%	-28%
SV-SE				223,052	-56%	28%
AR-SA				175,560	21%	19%

Table 5: Productivity and PED gains and losses between June 2018 and November 2019.

The higher the productivity, the better; the lower the PED, the better. In the above table a positive percentage for productivity (Prod) means that the hourly edited words increased by x%, while a negative value indicates productivity loss. Conversely, a negative PEDe value means that average number of edits fell over time (good), while a positive value means an increase in edits (bad).

For Good enough content, PEDe fell for our top nine languages. Likewise, productivity increased for these languages for High quality content. Productivity for Good enough content and PEDe for High quality content, however, have outliers.

Ideally, falling PEDe should result in rising productivity. While DE-DE exemplifies hoped-for results, there are obvious exceptions. Korean, for example, at the Good enough quality tier has PEDe falling by 48%, while productivity is dropping by 7%, when it should be rising in correlation.

The table below shows how the numbers above came about.

EN > DE	Good enough	High quality
Words	1,027,343	1,862,915
Prod Start	1200	1010
Prod Finish	1600	1150
Increase	33%	14%
PED Start	29	26
PED Finish	15	20
Reduction	48%	23%

Table 6: Productivity and PED gains [and losses] expressed in percentages.

The following two graphs provide bird's-eye views of trendlines for our top nine languages between January 2018 and November 2019. This timeframe spans the three MSFT Translator deployment phases of customized SMT, generic NMT and customized NMT. Please zoom in.

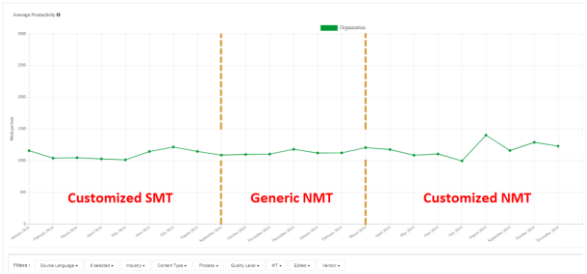

Figure 1: Productivity for 9 languages rising from 1050 words per hour to 1300.

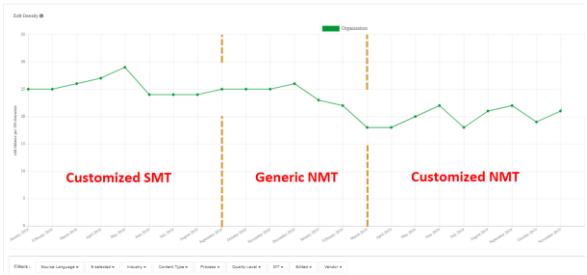

Figure 2: PEDe for 9 languages falling from 25 to 20 characters per 100 characters.

The trendline for PEDe appears to be more robust than for productivity. Also, within the discernable trends over 23 months, we see monthly ups and downs, suggesting that productivity is driven by multiple factors, not only MT quality.

The following graph compares productivity trendline and PEDe for EN-US to Dutch, a language pair for which we never customized NMT systems.

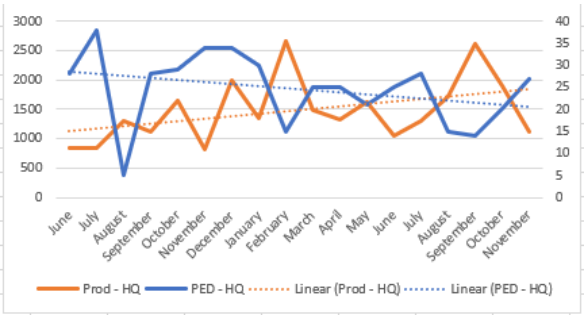

Figure 3: Well correlated trendlines: falling PED and rising productivity, going from generic SMT to generic NMT.

In summary, we can observe the overall benefits of customized NMT systems, especially when rolling up trendlines for all top nine languages. Looking at individual language pairs we can see exceptions. Pronounced discrepancies between PED and productivity we plan to review with our translation and MT technology providers.

6 Cut-off between MT and TM

In the following exercise, we wanted to find out if post-editing machine translations is faster than editing fuzzy matches. Depending on the language pair and NMT system, customized vs. generic, machine translated segments require fewer post-edits than fuzzy matches (Zaretskaya, 2019). Assuming that fewer post-edits mean shorter post-editing times, i.e., higher productivity, we should raise the MT-to-TM cut-off from 75% to x.

We looked at jobs machine translated against our customized NMT system from EN-US to DE-DE between March 2019 and 2020.

Match rate	Segments	Sentences per bracket	Bracket	Milliseconds
0	15,090	15,090	0	43,812
75	314			
76	304			
77	323			
78	319			
79	324	1,584	75-79	25,032
80	324			
81	328			
82	299			
83	305			
84	315	3,155	75-84	24,584
85	385			
86	424			
87	395			
88	379			
89	333			
90	362			
91	347			
92	355			
93	393			
94	404	3,777	85-94	20,968
95	405			
96	382			
97	670			
98	1,500	2,957	95-98	13,199
99	12,096	12,096	99	3,700
100	101,916	101,916	100	58
Grand Total	138,991			

Table 7: Distribution of 138,991 segments on leverage scale of 0 – 100, grouped by matching bands.

As expected, post-editing times diminish with increasing match rates. But it takes significantly longer to post-edit MT segments than 75%-84% fuzzies: 44 vs. 25 seconds. Increasing the MT-to-TM cut-off to 85% would drastically reduce productivity.

EN-US > DE-DE	Time in segs	Avg. secs	Total segs
0 - 10 nPEDi	Any	44	14,889
75-84 fuzzies	Any	25	3,155
0 - 10 nPEDi	0 - 120 secs	22	13,537
75 - 84 fuzzies	0 - 120 secs	17	3,044
0 - 10 nPEDi	0 - 60 secs	15	12,118
75 - 84 fuzzies	0 - 60 secs	12	2,848

Table 8: EN-US-to-DE-DE MT and fuzzy-match segments grouped by average active time. "Any" contains outliers.

Sorting segments by post-edit time, we noticed outliers that were active much longer than typically necessary for editing. The biggest outlier was active in the CAT tool for 16 minutes. When only considering segments active for 60 seconds – which means 81% of all MT segments - the editing time gap shrinks from 76% to 25%.

EN-US > FR-FR	Time in segs	Avg. secs	Total segs
0 - 10 nPEDi	Any	42	13,515
75-84 fuzzies	Any	26	2,866
0 - 10 nPEDi	0 - 120 secs	20	12,283
75 - 84 fuzzies	0 - 120 secs	15	2,724
0 - 10 nPEDi	0 - 60 secs	12	11,010
75 - 84 fuzzies	0 - 60 secs	11	2,561

Table 9: EN-US-to-FR-FR MT and fuzzy-match segments grouped by average active time. "Any" contains outliers.

For the EN-US-to-FR-FR language pair, the post-editing time gap between fuzzy matches and MT shrinks to 9% when only considering segments active for 60 seconds. This convergence likely applies to other well-performing MT language pairs as well: English to Brazilian Portuguese, Chinese, French, Italian and Spanish.

But as long as editing time for MT segments doesn't fall below editing time for fuzzy matches, raising the MT-to-TM cut-off would be counterproductive.

Let's see if the PED tells a different story.

EN > DE: w/o OL	MT	75-84	85-94	95-98	99
Segments	14,302	2,758	3,363	2,487	11,926
Avg aPEDi	23	23	16	12	3
Avg Words	14	10	14	12	7

Table 10: Average aPEDi for EN-US to DE-DE by match band, without outliers.

For EN-US to DE-DE, the average absolute PED for MT and the adjacent fuzzy match segments is the same, 23 characters. But, it takes 25% longer to edit the MT segments.

EN > FR: w/o OL	MT	75-84	85-94	95-98	99
Segments	13,180	2,675	2,708	1,836	14,769
Avg aPEDi	17	21	15	10	2
Avg Words	17	13	18	14	9

Table 11: Average aPEDi for EN-US to FR-FR by match band, without outliers.

For EN-US to FR-FR, the average PED for MT segments is 17 characters vs. 21 characters for 75%-84% fuzzy matches. While MT segments require 19% fewer post-edits, it takes 9% longer to edit them. It has been noted before, that post-edits and post-editing time, i.e., technical and temporal efforts do not necessarily correlate (Krings, 2001).

While we couldn't demonstrate that our MT segments can be edited faster than our fuzzies, we did gain a couple of useful insights. For one,

we need to optionally exclude outliers from our data to produce a richer picture of our MT productivity. About 10% of MT segments inflate both average post-editing time and PED noticeably.

EN > DE: with OL	MT	75-84	85-94	95-98	99
Segments	14,889	3,155	3,777	2,957	12,096
Avg aPEDi	28	75	80	62	6
Avg Words	15	11	15	12	8

Table 12: Average aPEDi for EN-US to DE-DE by match band, with outliers.

EN > FR: with OL	MT	75-84	85-94	95-98	99
Segments	13,515	2,866	2,958	2,180	14,881
Avg aPEDi	20	43	47	48	4
Avg Words	17	14	19	14	9

Table 13: Average aPEDi for EN-US to FR-FR by match band, with outliers.

The tables above contain outliers, segments with inline tags relatively easy to handle by linguists in the CAT tool, but with large character counts to the Levenshtein algorithm. In one sample, a segment with 7 words and 6 tags resulted in an aPEDi of 1015 characters. Going by the standard ratio of 1:5 for words to characters, the calculated aPEDi vastly overstates the human effort of placing a few tags and minor textual changes.

For two, 99% fuzzy matches deserve special consideration in SLAs, assuming they constitute a good portion of overall word count. In our 12-month sample, they account for 9% of total words, and require a fraction of editing time compared to other match bands. They ought to be broken out for dedicated costing.

7 Time spent on 0 nPEDi segments

In this section we discuss the time linguists spent on segments that required no post-editing. Ideally, CAT tools should flag these segments to linguists so that they can skip them.

The following table breaks down segments machine translated from EN-US into DE-DE between March 2019 and March 2020.

EN-US > DE-DE	0 nPEDi	1 nPEDi	0 - 10 nPEDi
HT + MT Segments	113,308	7,520	138,991
MT Segments	1,890	3,210	14,889
Segs with 0 Time in segment	1,389	219	1,778
Segs with Time in segment	501	2,991	13,111
Avg. time in segment *	8	19	50
Avg. segment length *	7	16	15

Table 14: Editing time for segments with 0 post-edits. Excluding segments with no time in segment (*).

Linguists didn't post-edit 1,890 (or 12%) of all MT segments. And the CAT tool didn't record post-editing time for 1389 segments of these unchanged segments, suggesting that the linguist had signed off on them unseen. We realized that the CAT tool allows linguists to sign off on segments w/o activating them. Because linguists can by-pass the time-tracker for unchanged segments, our hourly MT productivity is slightly overstated.

Of the unchanged segments (0 nPEDi), linguists did activate 501 for review. These segments were on average 7 words long and took linguists 8 seconds on average to conclude that no edits were needed.

In the 1 nPEDi bracket, linguists made minor changes, e.g., to correct compounds, punctuation, or word casing. For 219 segments the CAT tool recorded changes, but no time in segment. We found that search and replace operations register as PED, but not editing time. In the 1 nPEDi bracket, for the 2991 segments requiring editing time, the segment length goes up to 16 words, in line with the overall MT segment length of 15 words, yet the average time to edit is only 19 seconds versus 50 seconds for all MT segments.

To increase ROI from MT, we would need to achieve three things: for accurate productivity tracking, the CAT tool needs to optionally force linguists to activate segments for sign-off, even if no post edits are needed. To increase the modest 12% of MT segments that do not require post-editing, we need to improve MT output by re-training our NMT systems against the latest base model and by adjusting our Style Guides to make allowances for immaterial linguistic deviations.

If we manage to increase the percentage of segments that don't require post-editing, we need to find a way to flag these for linguists in the CAT tool via quality estimation models. Similar to how we opt to selectively skip review of repetitions or 100% matches, we may choose to skip review of low-risk MT segments.

8 Segments by nPEDi

To understand how segments are distributed on the nPEDi scale of 1 – 10, we looked at Good enough and High quality material, machine translated with our customized NMT system between March 2019 and February 2020.

The following diagram shows that most segments for FR-FR and DE-DE fall into the nPEDi range of 0 to 4. In line with the overall ranking of languages by productivity, FR-FR performs better than DE-DE, with more segments in the nPEDi range of 0 to 1.

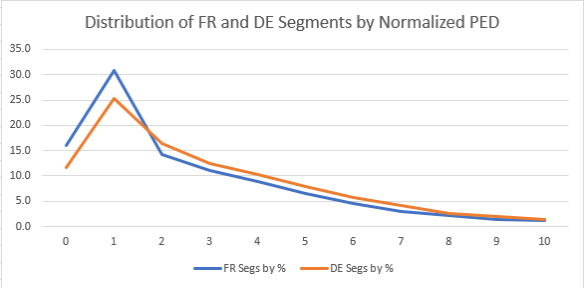

Figure 4: Distribution of segments on nPEDi scale of 0 to 10.

Since we omitted to track productivity for HT workflows, we need to go by the industry assumption that linguists are faster post-editing MT output than translating from scratch (HT) if the nPEDi is below 5. Going by this assumption, our customized NMT system boosts FR-FR productivity for 81% of the segments.

Norm. PED	Segs by #	FR Segs by %	Cummulative %
0	2,301	16.0	16.0
1	4,462	30.9	46.9
2	2,057	14.3	61.1
3	1,590	11.0	72.2
4	1,280	8.9	81.0
5	933	6.5	
6	669	4.6	
7	442	3.1	
8	304	2.1	
9	212	1.5	
10	175	1.2	
	14,425		

Table 15: nPEDi distribution for EN-US to FR-FR segments by percentages.

For DE-DE, our customized NMT system boosts productivity for 76% of the segments.

Norm. PED	Segs by #	DE Segs by %	Cummulative %
0	2,293	11.7	11.7
1	4,959	25.3	37.0
2	3,228	16.5	53.5
3	2,449	12.5	66.0
4	2,009	10.3	76.3
5	1,547	7.9	
6	1,112	5.7	
7	806	4.1	
8	504	2.6	
9	391	2.0	
10	292	1.5	
	19,590		

Table 16: nPEDi distribution for EN-US to DE-DE segments by percentages.

We are planning to analyze segments in the nPEDi range of 1 to 2 to understand if aligning styleguide requirements to MT capabilities or automated post-editing rules will elevate these low nPEDi to 0 nPEDi segments.

9 Correlating nPEDi and segment length

We approached the exercise of correlating nPEDi and segment length with the assumption that MT systems translate segments of a certain length best, segments that are not too short and not too long. Similar to linguists, MT systems may struggle with short segments for lack of context and with long segments because of complexity.

We tried to confirm this assumption with two different methods on segments machine translated with our customized NMT systems between March 2019 and February 2020.

In the first exercise, we simply expanded the nPEDi distribution table, by adding the total source language word count for each nPEDi mark and dividing it by the number of segments.

nPEDi	Segs	Words	FR W / S
0	2,301	22,821	10
1	4,462	69,311	16
2	2,057	30,723	15
3	1,590	23,312	15
4	1,280	18,483	14
5	933	12,673	14
6	669	8,056	12
7	442	4,242	10
8	304	2,730	9
9	212	1,511	7
10	175	950	5
	14,425	194,812	14

Table 17: Average segment length per nPEDi bracket.

nPEDi	Segs	Words	DE W / S
0	2,293	16,461	7
1	4,959	77,968	16
2	3,228	50,936	16
3	2,449	42,690	17
4	2,009	34,667	17
5	1,547	25,172	16
6	1,112	17,139	15
7	806	11,070	14
8	504	5,923	12
9	391	3,512	9
10	292	1,773	6
	19,590	287,311	15

Table 18: Average segment length per nPEDi bracket.

While short segments appear at both ends of the 1 – 10 nPEDi scale, 75% of them are in the 0 nPEDi bracket. This does mean that shorter strings machine translate more successfully.

In the next method, we grouped DE-DE segments by word length and, calculated the average aPEDi and nPEDi for each length.

Length	Segments	Avg. nPEDi	Avg. aPEDi
1	317	2.06	3
2	776	2.77	6
3	945	2.95	8
4	958	3.15	11
5	988	2.69	11
6	939	2.89	13
7	914	2.96	16
8	866	2.94	18
9	812	3.03	20
10	751	2.9	21
11	713	3.03	23
12	678	2.94	25
13	680	2.92	26
14	666	2.89	27
15	636	2.91	28
20	507	2.94	38
25	324	3.04	46
30	226	3.15	60
35	100	3.12	71
40	58	2.7	65
	12,854		

Table 19: Average nPEDi and aPEDi by segment length in words.

With the exception of segments 1, 2, 5 and 40 words long, the nPEDi hovers around 3 for segments of any length, even for the longer ones. We assume that the average nPEDi for longer segments doesn't increase noticeably because linguists may revisit these long and complicated segments in separate CAT sessions and therefore only record a portion of the actual post-edits.

The lighter (orange) line in the graph below illustrates the lower nPEDi for the very short segments and the otherwise relatively stable nPEDi:

Figure 5: nPEDi in characters by segment length in words.

Naturally, the aPEDi generally increases with segment length. The lines diverge because segment length is counted in words while aPEDi is counted in characters.

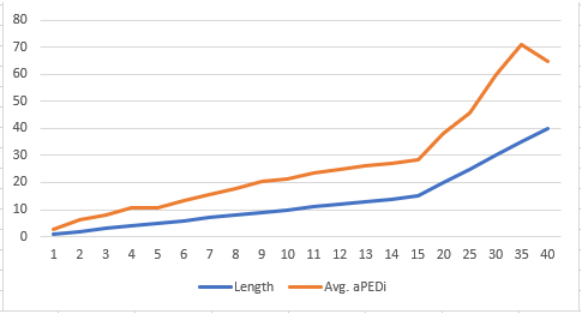

Figure 6: aPEDi in characters by segment length in words.

The first exercise clearly indicates that it is the majority of shorter segments that machine translate better than longer strings. A minority of shorter strings fails to translate well, probably for lack of context. The second exercise is inconclusive, but highlights one of the imperfections in our productivity tracking setup.

10 Future work

Users will benefit from several enhancements to GlobalLink to more accurately record post-editing efforts in this particular translation management system. Above all, the setup needs to capture all edits, whether performed in one or multiple CAT tool sessions. MT segments should only be included in the metrics if linguists signed off on them. To track editing time comprehensively, the CAT tool should optionally require activation also for segments that do not need post-edits. Batch changes that do not trigger the time tracker should be accounted for somehow. Tracking editing effort for HT segments is essential to establish a baseline. The comprehensive audit trail capabilities of GlobalLink allowed us to pinpoint these opportunities for improvement.

We hope that TAUS will use our findings to develop certification test plans for tools integrators to ensure that productivity metrics are consistently calculated across organizations using different CAT tools. Also, the TAUS DQF Dashboard should allow users to exclude outliers for an alternate productivity view.

For full access to our productivity data currently stored in the TAUS DQF Dashboard, Dell needs to integrate its BI tools.

11 Conclusion

Tracking MT productivity at scale needs to become an integral capability in the translation industry to be available regardless of which tools and services providers we partner with.

Even though our productivity metrics are after the fact, they are a statistically robust addition to small-scale human evaluations, BLEU scores and emerging risk calculation models. Together, these MT quality assurance methods help us focus our continuous improvement efforts.

Our numbers show that we are on the right track: productivity is steadily rising and post-edits falling. Our challenge will be to turn the many segments requiring few post-edits to ones that require none and to flag these segments in the CAT tool so that linguists can skip them.

We like to think that the imperfections we discovered in our setup balance each other out as some inflate and some deflate our productivity numbers. We are also reminded that within MTPE jobs, we apply machine translations to new words only, about 10% of total word count. The remaining 90% are leveraged from translation memories. While MT is an important productivity aid, it is not the only one in a linguist's tool chest.

Overall, 70% of our production translation jobs use MT to pre-translate new words. We will expand MT usage by starting to pre-translate software as well. The biggest expansion of MT usage at Dell, however, occurs somewhere else. To operate within a global enterprise, many of our colleagues produce raw MT in self-service mode. And data scientists machine translate vast amounts of data into English for processing by BI engines. This last use case dwarfs all others by volume. Closely monitoring machine translation quality in our human-assisted production workflows will benefit Dell's two use cases of unedited MT output as well.

Acknowledgements

The following individuals and organizations were instrumental in creating an environment to harvest MT metrics automatically for Dell EMC: Nancy Anderson, head of the EMC translation team at the time supported the proposal to take translations "online". She negotiated with our LSPs the necessary process and tools concessions. Keith Brazil and his team at Translations.com optimized GlobalLink as a collaborative platform for a multi-vendor supply chain. Jaap van der Meer proposed an integration with the TAUS DQF Dashboard. TAUS and

Translations.com then worked together to connect systems to calculate and record MT productivity metrics. Last but not least, our MLVs, their numerous SLVs and the many linguists agreed to trade their preferred CAT tools for a common technology platform.

References

Levenshtein, Vladimir Iosifovich. 1966. *Binary Codes Capable of Correcting Deletions, Insertions and Reversals*. Soviet Physics Doklady.

Zaretskaya, Anna. 2019. *Raising the TM Threshold in Neural MT Post-Editing: a Case-Study on Two Datasets.* Proceedings of MT Summit XVII, volume 2. 213-218.

Krings, Hans P. 2001. *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes.* Kent State University Press, Kent, Ohio.

Koponen, Maarit. 2016. *Is Machine Translation Postediting Worth the Effort? A Survey of Research into Post-editing and Effort.* The Journal of Specialised Translation.