DATA DESCRIPTION AND SUPPLEMENTARY DATA

MT Quality Evaluation

Data description Results of the evaluation carried out by human evaluators based on the MQM framework

"GT" = Google Translate
"DL" = Deep
"SY".= Systran Translate

01-total errors, sd, av, weightedav.xlsx
For all the sheets (GT, DL, SY), Column A indicates the number of texts in a crescent order from 1 to 52. Row 2 represents all the error types and severity attributed by the evaluators.
Column AC reports the error sum per text.
Row 55 represents the total number of errors reported for all the error types.
Column AC at row 55 reports the total number of error Google was attributed.
Column AC row 56 reports the average of errors made by Google.
Column AC row 57 reports the standard deviation of the three MT tools.
Column AC row 58 reports the weighted average.
The sheet named "Graphs" reports the graph version of the 4 main results, respectively total number of errors, standard deviation, average and weighted average.

02-error severity.xlsx
It reports the results of the errors from the perspective of severity.

03-error categories.xlsx
This figure reports the 4 error categories analysed by the manual evaluation, namely Accuracy, Fluency, Terminology and Other.
Some error categories group together more than one error type. This is the case for Accuracy and Fluency.
Accuracy is composed by Mistranslation, Addition/Omission, Untranslated/DNT.
Fluency is composed by Grammar/Syntax, Register, Inconsistency, Spelling.

04-error types.xlsx
This represents all the error categories.

Average time of post-editing
Their average times are summarised in the table below.

	Text 15	Text 22	Text 2	Text 27
Av. time DL	**21min28sec**	**16min16sec**	**21min38sec**	**34min12sec**
Av. time SY	22min51sec	16min22sec	23min36sec	44min44sec
Av. time GT	28min10sec	21min54sec	31min50sec	47min02sec

Additional Analysis

05-readability.xlsx
Reports for each text, for each MT engine, the text readability score and the average error

06-length.xlsx
Reports for each text, for each MT engine, the text length and the average error

SUPPLEMENTARY DATA

Results of the MT quality evaluation performed by using automated metrics

BERT SCORE
BLEU
TER
COMET

Table 1

MEAN	BERT-SCORE MEAN	BERT-SCORE SD	BLEU	BLEU SD	TER	TER SD	COMET-QE	COMET-QE SD
GT	**0.9115**	0.0168	**39.5180**	5.4260	**44.3536**	5.8060	0.5619	0.1320
DL	0.9106	0.0175	39.0896	5.3724	45.8873	5.5374	**0.5967**	0.1220
SY	0.9076	0.0165	37.5965	5.6701	46.6953	5.8333	0.5380	0.1366

Table 1 reports the mean of every method for each translation engine and its standard deviation.

Table 2

TEXT	MT	BERT-SCORE	TER	BLEU	COMET-QE	WINNER
44	GT	**0.906092**	**38.78**	**43.08**	0.69	1
44	DL	0.903147	42.32	41.29	**0.73**	0
44	SY	0.898369	43.41	40.61	0.56	0

Table 2 is a possible way to count the corpus data. We count how many times the translation systems “won” in their scores in total. This analysis can be performed in different ways. Table 2 is an example that explains the first method. This is the evaluation made for Text 44. Since 3 methods out of 4 set out that Google did best, according to the metrics ’scores, for Text 44 Google was appointed “winner”.

Table 3

TEXT	MT	BERT-SCORE	TER	BLEU	COMET-QE	WINNER
27	DL	**0.937488**	44.85	38.24	**0.66**	1
27	GT	0.923904	**44.29**	**39.60**	0.62	1
27	SY	0.921568	45.96	38.52	0.63	0

Table 3 applies the same method as table 2 but in a less clear case. In this case, Text 27 reports a doubtful evaluation. Since BERT and COMET find out that DeepL was the best translation system, and TER and BLEU set out that Google did best, both methods are evaluated as "winners".

Table 4

	BERT-SCORE	TER	BLEU	COMET-QE
DL	0.943842; Text **46**	30.97; Text 39	51.95; Text **46**	0.82; Text 16
GT	0.943019; Text 47	32.74; Text **49**	50.27; Text **49**	0.81; Text 46
SY	0.936107; Text **16**	35.34; Text **46**	46.95; Text **46**	0.76; Text **16**

Table 4 shows the best translated texts of the corpus and by which machine translation tool. In order to understand which was the best translation tool, we pick the highest points assigned by each method (for TER it will be the lowest). The table below shows the points scored by the best translation systems and reports the text that is considered as the "best" translated for each tool and method.

Table 5

TEXT	MT	BERT-SCORE	TER	BLEU	COMET-QE	WINNER
46	**DL**	**0.943842**	**30.97**	**51.95**	0.79	1
46	GT	0.938840	33.52	49.04	**0.81**	0
46	SY	0.932996	35.34	46.95	0.75	0

Here we analyse which text is considered "best" for each translation system. For DeepL it is clear, as said above, that it performed best in Text 46. Table 5 below compares of the scores DeepL and the other translation tools made for this same text.

Table 6

TEXT	MT	BERT-SCORE	TER	BLEU	COMET-QE	WINNER
49	DL	**0.941025**	35.84	47.13	**0.75**	1
49	GT	0.938756	**32.74**	**50.27**	0.68	1
49	SY	0.932126	38.50	41.48	0.70	0

Table 6 reports the results for Google Translate, which shows that the best accredited translation is Text 49 for TER and BLEU.

Table 7

TEXT	MT	BERT-SCORE	TER	BLEU	COMET-QE	WINNER
16	SY	**0.936107**	42.58	39.25	0.76	0
16	DL	0.935215	41.29	40.00	**0.82**	0
16	GT	0.933136	**40.97**	**42.14**	0.76	1

This final table shows that Systran has two best translations, because for TER and BLEU it translated Text 46 better than the other texts, while for BERT and COMET it translated Text 16 better.