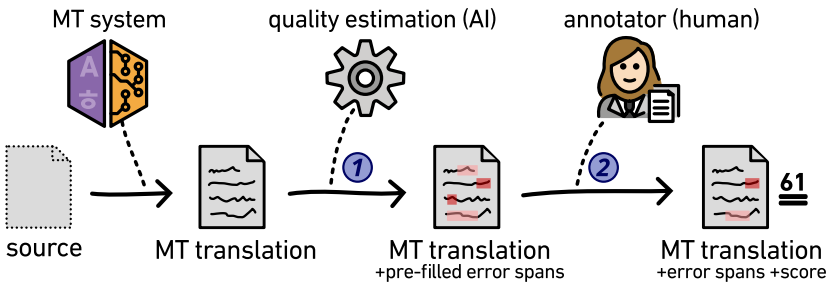
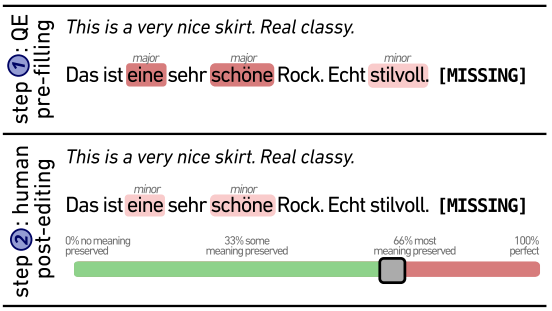
Even Better Human Evaluation of Machine Translation

AI-Assisted Human Evaluation of Machine Translation; Vilém Zouhar,[=] Tom Kocmi,[=] Mrinmaya Sachan





Faster, more annotated errors & higher agreement

	ESA	ESA^{AI}	
Annotated errors	0.45	1.63	pre-annotations make annotators notice
Average score	81.8	76.7	← more mistakes
Time/segment	58s	52s	← a bit faster
Time/segment/error	71s	31s	← much faster

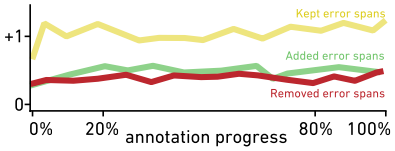
	Intra Agreement*		Inter Agreement	
	ESA	ESA^{AI}	ESA	ESA^{AI}
Score from spans	0.282	0.489	0.327	0.671
Direct scores	0.222	0.486	0.376	0.533

final score is based on human only → alleviates bias

Low annotator automation bias

Annotators make similar types of annotations at the beginning as at the end → No learned carelessness.

After attention checks, where QE is incorrect (does not detect errors), annotators trust it less and accept fewer QE suggestions (84%→73%).

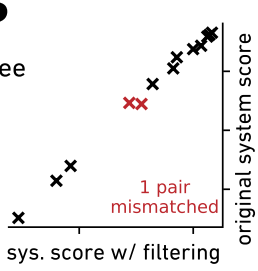


Automated pre-selection of evaluation-worthiness

When QE pre-annotates no errors, 90% annotators agree there are no errors. → Why human-annotated these?

With no-error spans removed (25% of budget), system ranking is almost the same!

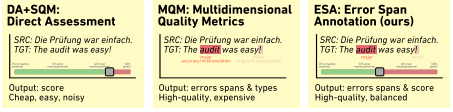
Check out subset2evaluate: How to Select Datapoints for Efficient Human Evaluation of NLG Models? (2025)



ESA: Efficient human evaluation of machine translation

Error Span Annotation: A Balanced Approach for Human Evaluation of Machine Translation

Tom Kocmi,[=] Vilém Zouhar,[=] Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, Mariya Shmatova

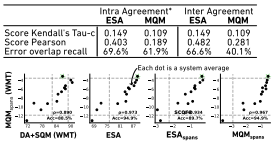


1.44x faster than MQM without need for experts

Our annotation campaign: 200 WMT23 English→German segments annotated by translators:
1) expert MQM annotators (49s/segment); and 2) ESA annotators (34s/segment)

Better quality than DA+SQM, comparable to MQM

	Intra Agreement*		Inter Agreement	
	ESA	MQM	ESA	MQM
Score Kendall's Tau-c	0.149	0.109	0.149	0.109
Score Pearson	0.403	0.189	0.482	0.281
Error overlap recall	69.6%	61.9%	66.6%	40.1%

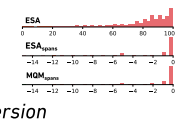


ESA replaced DA+SQM as the human evaluation protocol for the General MT Shared Task (News) at WMT24.

ESA gives continuous scores

MQM score is a weighted number of errors, with exceptions:
MQM$_{spans}$ = -1× #minor -5× #major
Majority of segments thus have score of 0, some -1, -2, or -5.

MQM scores segments discretely with low noise.
DA scores segments continuously but is noisy.
ESA scores segments continuously with low noise.



human-only version