A Appendix A

Table 6: Segment-level Kendall's Tau (τ) correlations on language pairs not involving English for the WMT19 Metrics DARR corpus. COMET-RANK (base) scores are to be replaced with results of the large model.

	de-cs	**de-fr**	**fr-de**	
Nº Tuples	23194	4862	1369	**avg.**
BLEU	0.222	0.226	0.173	0.207
CHRF	0.341	0.287	0.274	0.301
BERTSCORE (F1)	0.356	0.330	0.277	0.321
PRISM	0.452	0.443	**0.421**	0.439
COMET-MQM (large)	0.413	0.422	0.327	0.387
COMET-HTER (large)	0.425	0.449	0.381	0.418
COMET-DA (large)	**0.471**	**0.469**	**0.420**	**0.453**
COMET-RANK (base)	0.389	0.444	0.331	0.388

Table 7: System-level Pearson correlation (r) for the from-English language pairs from WMT19 DA corpus. DARR Ranker (base) scores are to be replaced with results of the large model.

	en-cs	**en-de**	**en-fi**	**en-gu**	**en-kk**	**en-lt**	**en-ru**	**en-zh**	
Nº Systems	11	22	12	11	10	12	12	12	**avg.**
BLEU	**0.988**	0.952	0.978	0.780	0.864	0.979	0.973	0.762	0.910
CHRF	**0.986**	0.983	**0.988**	0.839	0.969	0.964	0.979	0.822	0.941
BERTSCORE (F1)	0.983	0.990	0.969	0.907	0.983	0.972	**0.989**	0.927	0.965
PRISM	0.964	0.987	0.947	-	0.978	0.929	0.914	0.900	0.946
COMET-MQM (large)	0.943	0.968	0.949	0.946	0.979	**0.985**	0.966	0.958	0.962
COMET-HTER (large)	0.948	**0.991**	0.959	0.948	0.965	**0.982**	0.973	0.943	0.964
COMET-DA (large)	0.964	**0.995**	0.969	**0.964**	**0.989**	**0.982**	**0.987**	**0.969**	**0.977**
COMET-RANK (base)	0.943	0.937	0.914	0.817	0.963	0.973	0.861	0.942	0.919

Table 8: System-level Pearson correlation (r) for the into-English language pairs from WMT19 DA corpus. DARR Ranker (base) scores are to be replaced with results of the large model.

	de-en	**fi-en**	**gu-en**	**kk-en**	**lt-en**	**ru-en**	**zh-en**	
Nº Systems	16	11	9	7	11	13	15	**avg.**
BLEU	0.879	0.984	0.975	0.959	0.969	0.840	0.895	0.929
CHRF	0.916	**0.988**	0.967	0.982	0.938	0.942	0.952	0.955
BERTSCORE (F1)	0.949	0.984	**0.990**	**0.995**	0.961	0.901	0.982	0.966
BLEURT (large-512)	0.939	0.984	0.989	0.989	**0.992**	**0.980**	**0.994**	**0.981**
PRISM	**0.954**	0.981	**0.992**	**0.992**	**0.994**	0.905	**0.992**	0.973
COMET-MQM (large)	0.926	0.974	0.972	0.971	0.986	0.889	0.959	0.954
COMET-HTER (large)	0.918	0.953	0.958	0.951	0.983	0.924	0.978	0.952
COMET-DA (large)	0.946	0.983	**0.993**	**0.996**	**0.993**	0.970	**0.993**	**0.982**
COMET-RANK (base)	0.922	0.981	0.963	0.932	0.987	0.674	0.967	0.918

Table 9: System-level Pearson correlation (r) for language pairs not involving English from WMT19 DA corpus.

	de-cs	**de-fr**	**fr-de**	
N^o Systems	9	11	10	**avg.**
BLEU	0.936	0.934	0.869	0.913
CHRF	**0.994**	0.933	0.908	0.945
BERTSCORE (F1)	0.988	0.953	0.942	0.961
PRISM	0.988	0.924	0.922	0.945
COMET-MQM (large)	0.936	0.950	0.885	0.924
COMET-HTER (large)	0.951	0.901	0.924	0.925
COMET-DA (large)	0.973	**0.972**	**0.954**	**0.966**
COMET-RANK (base)	0.819	0.941	0.927	0.896

Table 10: Document-level Pearson correlation (r) for micro average and macro average for English-to-German and English-to-Czech.

	en-cs		**en-de**	
	Micro-avg.	Macro-avg.	Micro-avg.	Macro-avg.
COMET-DA (large)	**0.667**	0.660	0.528	**0.529**
COMET-MQM (large)	0.638	**0.639**	0.516	**0.519**
COMET-HTER (large)	**0.655**	0.650	**0.558**	0.552
	0.653	0.649	**0.534**	0.533