

	fr-en (6268 pairs)	de-en (6382 pairs)	es-en (4106 pairs)	cz-en (2251 pairs)	hu-en (2193 pairs)	xx-en (1949 pairs)	Overall (23149 pairs)
ulc	.64	.64	.61	.63	.60	.63	.63
rte (absolute)	.64	.62	.61	.62	.65	.62	.62
maxsim	.63	.63	.61	.60	.62	.61	.62
wcd6p4er	.63	.62	.61	.61	.55	.57	.61
wpF	.63	.59	.59	.61	.56	.60	.60
terp	.62	.59	.60	.60	.55	.57	.60
bleusp	.62	.60	.57	.59	.55	.58	.59
bleusp4114	.61	.60	.57	.59	.55	.58	.59
ter	.55	.53	.52	.50	.45	.50	.52
rte (pairwise)	.44	.45	.51	.59	.64	.63	.51
wpbleu	.53	.49	.51	.49	.44	.57	.51
meteor-0.7	.51	.52	.49	.48	.48	.44	.50
meteor-0.6	.51	.52	.49	.47	.48	.44	.50
meteor-ranking	.51	.52	.49	.47	.48	.44	.50

Table 10: Sentence-level consistency of the automatic metrics with human judgments for translations into English. Italicized numbers do not beat the random-choice baseline. **(This table was corrected after publication.)**

	fr-en (6268 pairs)	de-en (6382 pairs)	es-en (4106 pairs)	cz-en (2251 pairs)	hu-en (2193 pairs)	Overall (21200 pairs)
Oracle	.61	.63	.59	.61	.67	.62
ulc	.61	.62	.58	.61	.59	.60
rte (absolute)	.60	.61	.59	.57	.65	.61
maxsim	.61	.62	.59	.57	.61	.60
wcd6p4er	.61	.59	.57	.55	.44	.57
wpF	.60	.59	.57	.61	.46	.58
terp	.60	.61	.59	.57	.56	.59
bleusp	.61	.59	.56	.55	.48	.57
bleusp4114	.61	.59	.56	.55	.46	.57
<i>ter</i>	.60	.59	.57	.55	.51	.58
<i>rte (pairwise)</i>	.56	.61	.57	.59	.64	.59
<i>wpbleu</i>	.60	.59	.57	.57	.43	.57
<i>meteor-0.7</i>	.61	.61	.58	.57	.55	.59
<i>meteor-0.6</i>	.61	.61	.58	.57	.60	.60
<i>meteor-rank</i>	.61	.61	.59	.57	.61	.60

Table 12: Consistency of the automatic metrics when their system-level ranks are treated as sentence-level scores. The scores in red italics indicate cases where the system-level ranks outperform a metric’s sentence-level ranks.

	en-fr (2967 pairs)	en-de (6563 pairs)	en-es (3249 pairs)	en-cz (11242 pairs)	Overall (24021 pairs)
wcd6p4er	.67	.58	.61	.59	.60
bleusp	.65	.56	.60	.56	.58
bleusp4114	.65	.56	.60	.56	.58
ter	.58	.50	.52	.44	.49
terp	.62	.50	.54	.31	.43
wpF	.66	.60	.61	n/a	.61
wpbleu	.60	.47	.49	n/a	.51

Table 11: Sentence-level consistency of the automatic metrics with human judgments for translations out of English. Italicized numbers do not beat the random-choice baseline. **(This table was corrected after publication.)**

	en-fr (2967 pairs)	en-de (6563 pairs)	en-es (3249 pairs)	en-cz (11242 pairs)	Overall (24021 pairs)
Oracle	.62	.59	.63	.60	.60
wcd6p4er	.62	.46	.58	.50	.52
bleusp	.62	.48	.59	.50	.52
bleusp4114	.63	.48	.59	.50	.52
<i>ter</i>	.61	.51	.58	.50	.53
<i>terp</i>	.62	.50	.59	.53	.54
wpF	.63	.50	.59	n/a	.55
<i>wpbleu</i>	.63	.51	.60	n/a	.56

Table 13: Consistency of the automatic metrics when their system-level ranks are treated as sentence-level scores. The scores in red italics indicate cases where the system-level ranks outperform a metric’s sentence-level ranks.