

	en-de (13 systems)	en-fr (16 systems)	en-es (11 systems)	en-cz (5 systems)	Average
terp	.03	-.89	-.58	<b>-.4</b>	<b>-.46</b>
ter	-.03	-.78	-.5	-.1	-.35
bleusp4114	-.3	.88	.51	.1	.3
bleusp	-.3	.87	.51	.1	.29
bleu	-.43	.87	.36	.3	.27
bleu (cased)	-.45	.87	.35	.3	.27
bleu-ter/2	-.37	.87	.44	.1	.26
wcd6p4er	.54	-.89	-.45	-.1	-.22
nist (cased)	-.47	.84	.35	.1	.2
nist	-.52	.87	.23	.1	.17
wpF	-.06	.9	.58	<i>n/a</i>	<i>n/a</i>
wpbleu	<b>.07</b>	<b>.92</b>	<b>.63</b>	<i>n/a</i>	<i>n/a</i>

Table 8: The system-level correlation of the automatic evaluation metrics with the human judgments for translation out of English.