The Effect of Translationese in Machine Translation Test Sets Supplementary Material

Mike Zhang
Information Science Programme
University of Groningen
The Netherlands
j.j.zhang.1@student.rug.nl

Antonio Toral
Center for Language and Cognition
University of Groningen
The Netherlands
a.toral.ruiz@rug.nl

A Supplemental Material

These are the supplementary tables for the paper "The Effect of Translationese in Machine Translation Test Sets". Provided are the remaining 16 tables of each language direction. These tables are of the same structure as Table 4 in the paper.

English→Chinese

	#	SYSTEM	RAW.WMT	Z.WMT	#	↑↓	SYSTEM	RAW.ORG	Z.ORG	#	↑↓	SYSTEM	RAW.TRS	Z.TRS
wmt17	1	SogouKnowing-nmt	73.2	0.208	1	–	SogouKnowing-nmt	69.1	0.063	1	$2^{\uparrow}$	xmunmt	78.2	0.396
		uedin-nmt	72.5	0.178		–	uedin-nmt	67.8	0.015		$1^{\downarrow}$	SogouKnowing-nmt	77.3	0.352
		xmunmt	72.0	0.165		–	xmunmt	66.0	-0.059		$1^{\downarrow}$	uedin-nmt	77.4	0.349
	4	online-B	69.8	0.065		$2^{\uparrow}$	CASICT-cons	63.6	-0.134		$1^{\uparrow}$	jhu-nmt	75.4	0.278
		jhu-nmt	69.5	0.056		$1^{\downarrow}$	online-B	64.7	-0.142		$1^{\downarrow}$	online-B	74.8	0.271
		CASICT-cons	68.5	0.035		$1^{\downarrow}$	jhu-nmt	63.3	-0.177		$1^{\uparrow}$	online-A	74.0	0.223
		online-A	68.2	0.010		–	online-A	62.5	-0.195		$1^{\downarrow}$	CASICT-cons	73.3	0.202
	8	Oregon-State-Uni-S	64.8	-0.111	8	–	Oregon-State-Uni-S	59.1	-0.338	8	–	Oregon-State-Uni-S	70.7	0.121
	9	UU-HNMT	59.2	-0.300	9	–	UU-HNMT	54.4	-0.499	9	–	UU-HNMT	64.5	-0.083
	10	online-G	55.9	-0.438	10	–	online-G	52.4	-0.599	10	–	online-G	59.4	-0.277
	11	online-F	53.1	-0.504		–	online-F	48.4	-0.668		–	online-F	57.7	-0.343
wmt18	1	Tencent-ensemble	80.7	0.219	1	–	Tencent-ensemble	76.7	0.062	1	–	Tencent-ensemble	83.0	0.314
		Unisound	80.3	0.206		–	Unisound	76.1	0.046		–	Unisound	82.9	0.301
		GTCOM-Primary	80.5	0.199		$2^{\uparrow}$	Alibaba-General-A	74.9	0.040		–	GTCOM-Primary	83.2	0.301
		Alibaba-ensemble	79.7	0.185		$3^{\uparrow}$	Alibaba-General-B	74.6	0.024		–	Alibaba-ensemble	82.0	0.281
		Alibaba-General-A	79.2	0.173		$2^{\downarrow}$	GTCOM-Primary	75.9	0.021		$1^{\uparrow}$	online-B	81.9	0.261
		online-B	79.5	0.166		$2^{\downarrow}$	Alibaba-ensemble	75.7	0.021		$1^{\downarrow}$	Alibaba-General-A	81.7	0.252
		Alibaba-General-B	79.0	0.165		$1^{\downarrow}$	online-B	75.6	0.011		–	Alibaba-General-B	81.6	0.249
	8	UMD	78.1	0.094		$1^{\uparrow}$	NICT	74.2	-0.050	8	–	UMD	81.3	0.209
		NICT	77.5	0.082		$1^{\downarrow}$	UMD	72.8	-0.101		$1^{\uparrow}$	online-Y	79.8	0.180
		online-Y	77.1	0.069		–	online-Y	72.7	-0.109		$1^{\uparrow}$	online-A	79.2	0.179
		online-A	75.5	0.037	11	–	online-A	69.3	-0.207		$2^{\downarrow}$	NICT	79.6	0.161
	12	uedin	70.7	-0.202	12	–	uedin	65.5	-0.473	12	–	uedin	73.9	-0.037
	13	online-F	63.3	-0.419	13	–	online-F	58.7	-0.607	13	$1^{\uparrow}$	online-G	65.6	-0.307
		online-G	63.4	-0.435		–	online-G	59.7	-0.647		$1^{\downarrow}$	online-F	66.0	-0.309

Table 1: Results of the English→Chinese language direction with WMT, ORG, and TRS. Systems are ordered by standardized mean DA score. If a system does not contain a rank, it indicates that it shares the same cluster as the system above it. Clusters are obtained according to Wilcoxon rank-sum test at p-level $p \leq 0.05$. Indicated in the [↑↓] column are the changes in absolute ranking (i.e. how many positions it goes up or down).

Czech→English

	#	SYSTEM	RAW.WMT	Z.WMT	#	↑↓	SYSTEM	RAW.ORG	Z.ORG	#	↑↓	SYSTEM	RAW.TRS	Z.TRS
wmt16	1	uedin-nmt	75.4	0.207	1	–	uedin-nmt	69.6	-0.010	1	–	uedin-nmt	81.1	0.421
	2	jhu-pbmt	72.6	0.101	2	–	jhu-pbmt	67.7	-0.073	2	–	jhu-pbmt	77.5	0.275
	3	online-B	70.8	0.051	3	–	online-B	66.1	-0.124		–	online-B	75.5	0.224
	4	online-A	69.5	0.000		–	online-A	64.8	-0.169	4	$1^{\uparrow}$	PJATK	75.3	0.197
		PJATK	69.0	-0.024	5	–	PJATK	62.7	-0.245		$1^{\downarrow}$	online-A	74.0	0.165
	6	cu-mergedtrees	55.8	-0.503	6	–	cu-mergedtrees	53.2	-0.599	6	–	cu-mergedtrees	58.4	-0.406
wmt17	1	uedin-nmt	74.6	0.181	1	–	uedin-nmt	70.3	0.018	1	–	uedin-nmt	78.8	0.343
	2	online-B	71.9	0.068	2	–	online-B	68.8	-0.049	2	–	online-B	74.9	0.185
	3	online-A	68.3	-0.068	3	–	online-A	64.7	-0.193	3	–	online-A	71.8	0.057
	4	PJATK	62.7	-0.268	4	–	PJATK	57.5	-0.462	4	–	PJATK	67.9	-0.074
wmt18	1	CUNI-Transformer	71.8	0.298	1	–	CUNI-Transformer	70.2	0.254	1	–	CUNI-Transformer	73.4	0.341
	2	uedin	67.9	0.165	2	–	uedin	65.9	0.104	2	–	uedin	70.0	0.225
	3	online-B	66.6	0.115		–	online-B	65.9	0.102	3	–	online-B	67.3	0.127
	4	online-A	62.1	-0.023	4	–	online-A	60.9	-0.051	4	–	online-A	63.2	0.004
	5	online-G	57.5	-0.183	5	–	online-G	55.5	-0.246	5	–	online-G	59.5	-0.120

Table 2: Results of the Czech→English language direction with WMT, ORG, and TRS. Systems are ordered by standardized mean DA score. If a system does not contain a rank, it indicates that it shares the same cluster as the system above it. Clusters are obtained according to Wilcoxon rank-sum test at p-level $p \leq 0.05$. Indicated in the [↑↓] column are the changes in absolute ranking (i.e. how many positions it goes up or down).

English→Czech

	#	SYSTEM	RAW.WMT	Z.WMT	#	↑↓	SYSTEM	RAW.ORG	Z.ORG	#	↑↓	SYSTEM	RAW.TRS	Z.TRS
wmt17	1	uedin-nmt	62.0	0.308	1	–	uedin-nmt	56.2	0.126	1	–	uedin-nmt	69.4	0.544
	2	online-B	59.7	0.240		–	online-B	55.1	0.090	2	–	online-B	64.8	0.405
	3	limsi-factored-norm	55.9	0.111	3	–	limsi-factored-norm	50.2	-0.074		–	limsi-factored-norm	63.5	0.354
		LIUM-FNMT	55.2	0.102		–	LIUM-FNMT	49.6	-0.076		$1^{\uparrow}$	LIUM-NMT	61.8	0.299
		LIUM-NMT	55.2	0.090		–	LIUM-NMT	49.5	-0.086		$1^{\downarrow}$	LIUM-FNMT	61.5	0.299
		CU-Chimera	54.1	0.050		–	CU-Chimera	48.2	-0.143		$1^{\uparrow}$	online-A	61.1	0.282
		online-A	53.3	0.029	7	–	online-A	45.2	-0.233		$1^{\downarrow}$	CU-Chimera	61.1	0.278
	8	TT-ufal_8gb	44.9	-0.236	8	–	TT-ufal_8gb	40.5	-0.380	8	–	TT-ufal_8gb	49.9	-0.075
	9	TT-afrl_4gb	42.2	-0.315	9	$1^{\uparrow}$	PJATK	36.9	-0.479		–	TT-afrl_4gb	48.0	-0.147
		PJATK	41.9	-0.327		$1^{\downarrow}$	TT-afrl_4gb	36.2	-0.491		$1^{\uparrow}$	TT-baseline_8gb	47.2	-0.169
		TT-baseline_8gb	40.7	-0.373		$1^{\uparrow}$	TT-afrl_8gb	35.1	-0.556		$1^{\downarrow}$	PJATK	46.9	-0.174
		TT-afrl_8gb	40.5	-0.376		$1^{\downarrow}$	TT-baseline_8gb	34.5	-0.565		–	TT-afrl_8gb	46.3	-0.184
	13	TT-ufal_4gb	36.5	-0.486		$1^{\uparrow}$	TT-denisov_4gb	33.1	-0.598	13	–	TT-ufal_4gb	42.2	-0.316
		TT-denisov_4gb	36.6	-0.493		$1^{\downarrow}$	TT-ufal_4gb	31.0	-0.647		–	TT-denisov_4gb	40.2	-0.386
wmt18	1	CUNI-Transformer	67.2	0.594	1	–	CUNI-Transformer	60.6	0.397	1	–	CUNI-Transformer	74.4	0.814
	2	uedin	60.6	0.384	2	–	uedin	52.4	0.120	2	–	uedin	69.6	0.674
	3	online-B	52.1	0.101	3	–	online-B	45.4	-0.098	3	–	online-B	59.3	0.315
	4	online-A	46.0	-0.115	4	–	online-A	36.6	-0.405	4	–	online-A	56.2	0.202
	5	online-G	42.0	-0.246		–	online-G	34.9	-0.467	5	–	online-G	49.7	-0.004

Table 3: Results of the English→Czech language direction with WMT, ORG, and TRS. Systems are ordered by standardized mean DA score. If a system does not contain a rank, it indicates that it shares the same cluster as the system above it. Clusters are obtained according to Wilcoxon rank-sum test at p-level $p \leq 0.05$. Indicated in the [↑↓] column are the changes in absolute ranking (i.e. how many positions it goes up or down).

Estonian→English

	#	SYSTEM	RAW.WMT	Z.WMT	#	↑↓	SYSTEM	RAW.ORG	Z.ORG	#	↑↓	SYSTEM	RAW.TRS	Z.TRS
	1	tilde-nc-nmt	73.3	0.326	1	–	tilde-nc-nmt	69.3	0.208	1	–	tilde-nc-nmt	77.3	0.444
	2	NICT	71.1	0.238	2	–	NICT	66.2	0.108	2	–	NICT	75.8	0.366
		tilde-c-nmt	69.9	0.215		–	tilde-c-nmt	66.1	0.101		–	tilde-c-nmt	73.9	0.331
		M4t1ss	69.0	0.187		$1^{\uparrow}$	uedin	65.6	0.060		–	M4t1ss	73.5	0.316
		uedin	69.2	0.186		$1^{\downarrow}$	M4t1ss	64.4	0.058		$1^{\uparrow}$	tilde-c-nmt-comb	73.0	0.312
		tilde-c-nmt-comb	68.7	0.171		–	tilde-c-nmt-comb	64.3	0.031		$1^{\downarrow}$	uedin	72.7	0.307
wmt18	7	online-B	67.1	0.117		–	online-B	64.7	0.030	7	$1^{\uparrow}$	HY-NMT-et-en	70.7	0.227
		HY-NMT-et-en	66.4	0.106		$1^{\uparrow}$	talp-upc	62.5	-0.003		$1^{\uparrow}$	talp-upc	71.0	0.214
		talp-upc	66.8	0.106		$1^{\downarrow}$	HY-NMT-et-en	61.9	-0.018		$2^{\downarrow}$	online-B	69.6	0.206
	10	online-A	65.4	0.063		–	online-A	62.2	-0.036		–	online-A	68.5	0.160
		CUNI-Kocmi	64.0	0.007	11	–	CUNI-Kocmi	59.3	-0.137	11	–	CUNI-Kocmi	68.4	0.145
	12	neurotolge.ee	59.4	-0.117	12	–	neurotolge.ee	54.4	-0.260	12	–	neurotolge.ee	64.7	0.032
	13	online-G	52.7	-0.341	13	–	online-G	52.9	-0.342	13	–	online-G	52.6	-0.340
	14	UnsupTartu	34.6	-0.950	14	–	UnsupTartu	34.4	-0.959	14	–	UnsupTartu	34.8	-0.941

Table 4: Results of the Estonian→English language direction with WMT, ORG, and TRS. Systems are ordered by standardized mean DA score. If a system does not contain a rank, it indicates that it shares the same cluster as the system above it. Clusters are obtained according to Wilcoxon rank-sum test at p-level $p \leq 0.05$. Indicated in the [↑↓] column are the changes in absolute ranking (i.e. how many positions it goes up or down).

English→Estonian

	#	SYSTEM	RAW.WMT	Z.WMT	#	↑↓	SYSTEM	RAW.ORG	Z.ORG	#	↑↓	SYSTEM	RAW.TRS	Z.TRS
	1	tilde-nc-nmt	64.9	0.549	1	–	tilde-nc-nmt	60.8	0.416	1	–	tilde-nc-nmt	68.8	0.676
	2	NICT	62.1	0.453		–	NICT	59.2	0.346	2	–	NICT	64.9	0.558
		tilde-c-nmt	61.6	0.427		–	tilde-c-nmt	58.5	0.317		$1^{\uparrow}$	M4t1ss	64.9	0.545
		M4t1ss	61.2	0.418		–	M4t1ss	57.5	0.289		$1^{\downarrow}$	tilde-c-nmt	64.6	0.534
	5	Aalto	58.6	0.340		–	Aalto	55.1	0.213		$1^{\uparrow}$	HY-NMT-en-et	62.7	0.464
		HY-NMT-en-et	58.6	0.329		$1^{\uparrow}$	uedin	54.7	0.198		$1^{\downarrow}$	Aalto	62.1	0.463
wmt18		uedin	57.5	0.295		$1^{\downarrow}$	HY-NMT-en-et	54.5	0.190		–	uedin	60.2	0.390
	8	CUNI-Kocmi	55.5	0.216		–	CUNI-Kocmi	54.4	0.174	8	$1^{\uparrow}$	talp-upc	57.9	0.292
		talp-upc	54.6	0.181	9	–	talp-upc	51.2	0.068		$1^{\downarrow}$	CUNI-Kocmi	56.7	0.258
	10	online-B	52.1	0.097		–	online-B	49.1	-0.010		–	online-B	55.1	0.201
	11	neurotolge.ee	45.7	-0.132	11	–	neurotolge.ee	43.6	-0.201	11	$1^{\uparrow}$	online-A	48.4	-0.047
	12	online-A	43.8	-0.195	12	–	online-A	39.2	-0.347		$1^{\downarrow}$	neurotolge.ee	47.8	-0.064
	13	online-G	37.6	-0.406	13	–	online-G	34.7	-0.508	13	–	online-G	40.5	-0.305
	14	parfda	34.3	-0.520	14	–	parfda	31.8	-0.604	14	–	parfda	36.7	-0.437

Table 5: Results of the English→Estonian language direction with WMT, ORG, and TRS. Systems are ordered by standardized mean DA score. If a system does not contain a rank, it indicates that it shares the same cluster as the system above it. Clusters are obtained according to Wilcoxon rank-sum test at p-level $p \leq 0.05$. Indicated in the [↑↓] column are the changes in absolute ranking (i.e. how many positions it goes up or down).

Finnish→English

	#	SYSTEM	RAW.WMT	Z.WMT	#	↑↓	SYSTEM	RAW.ORG	Z.ORG	#	↑↓	SYSTEM	RAW.TRS	Z.TRS
wmt16	1	online-B	66.9	0.095	1	–	online-B	63.7	-0.005	1	$2^{\uparrow}$	online-G	69.9	0.220
		uedin-pbmt	66.3	0.087		–	uedin-pbmt	63.1	-0.034		–	uedin-pbmt	69.6	0.210
		online-G	66.4	0.084		–	online-G	62.9	-0.051		$1^{\uparrow}$	UH-opus	70.0	0.207
		UH-opus	65.9	0.065		–	UH-opus	61.8	-0.078		$3^{\downarrow}$	online-B	70.1	0.195
	5	PROMT-SMT	62.9	-0.037	5	–	PROMT-SMT	60.3	-0.136	5	–	PROMT-SMT	65.5	0.063
	6	uedin-syntax	61.5	-0.090		–	uedin-syntax	59.0	-0.174	6	–	uedin-syntax	64.0	-0.007
		UH-factored	61.2	-0.098		–	UH-factored	58.6	-0.180		–	UH-factored	63.7	-0.016
		online-A	60.6	-0.126		–	online-A	58.1	-0.208		–	online-A	63.2	-0.044
	9	jhu-pbmt	52.7	-0.391	9	–	jhu-pbmt	51.8	-0.425	9	–	jhu-pbmt	53.6	-0.357
wmt17	1	online-B	73.8	0.407	1	–	online-B	71.7	0.324	1	–	online-B	76.0	0.490
	2	online-G	67.5	0.220	2	–	online-G	63.8	0.086	2	–	online-G	71.2	0.358
	3	online-A	62.6	0.041	3	–	online-A	59.3	-0.066	3	–	online-A	66.0	0.151
	4	TALP-UPC	58.8	-0.095		–	TALP-UPC	58.2	-0.120	4	–	TALP-UPC	59.5	-0.069
	5	Hunter-MT	52.1	-0.316	5	–	Hunter-MT	49.3	-0.396	5	–	Hunter-MT	55.0	-0.237
	6	apertium	44.6	-0.559	6	–	apertium	42.1	-0.648	6	–	apertium	47.1	-0.469
wmt18	1	NICT	75.2	0.153	1	–	NICT	72.8	0.086	1	–	NICT	77.5	0.218
		HY-NMT-fi-en	74.4	0.128		–	HY-NMT-fi-en	72.2	0.044		–	HY-NMT-fi-en	76.7	0.216
		uedin	74.0	0.103		$3^{\uparrow}$	talp-upc	71.7	0.028		–	uedin	77.4	0.200
		CUNI-Kocmi	72.7	0.083		$3^{\uparrow}$	online-A	71.2	0.027		–	CUNI-Kocmi	76.2	0.187
		online-B	72.9	0.078		–	online-B	71.2	0.020		–	online-B	74.5	0.136
		talp-upc	71.9	0.047		$3^{\downarrow}$	uedin	70.7	0.008	6	–	talp-upc	72.1	0.066
		online-A	71.5	0.045		$3^{\downarrow}$	CUNI-Kocmi	69.2	-0.024		–	online-A	71.9	0.064
	8	online-G	66.1	-0.134	8	–	online-G	63.0	-0.250	8	–	online-G	69.2	-0.019
	9	JUCBNMT	58.9	-0.404	9	–	JUCBNMT	57.3	-0.480	9	–	JUCBNMT	60.6	-0.325

Table 6: Results of the Finnish→English language direction with WMT, ORG, and TRS. Systems are ordered by standardized mean DA score. If a system does not contain a rank, it indicates that it shares the same cluster as the system above it. Clusters are obtained according to Wilcoxon rank-sum test at p-level $p \leq 0.05$. Indicated in the [↑↓] column are the changes in absolute ranking (i.e. how many positions it goes up or down).

English→Finnish

	#	SYSTEM	RAW.WMT	Z.WMT	#	↑↓	SYSTEM	RAW.ORG	Z.ORG	#	↑↓	SYSTEM	RAW.TRS	Z.TRS
wmt17	1	online-B	59.6	0.378	1	–	online-B	54.5	0.209	1	–	online-B	65.2	0.561
		HY-HNMT	57.8	0.305	2	–	HY-HNMT	51.7	0.096		–	HY-HNMT	64.4	0.534
	3	online-G	51.6	0.090	3	–	online-G	48.5	-0.026	3	$2^{\uparrow}$	AaltoHnmtMulti	55.8	0.236
		jhu-nmt-lattice	51.3	0.060	4	–	jhu-nmt-lattice	46.4	-0.108		–	jhu-nmt-lattice	56.0	0.220
		AaltoHnmtMulti	49.3	-0.004		–	AaltoHnmtMulti	43.8	-0.208		$2^{\downarrow}$	online-G	54.6	0.201
	6	AaltoHnmtFlatcat	46.4	-0.102		–	AaltoHnmtFlatcat	42.7	-0.245	6	$2^{\uparrow}$	HY-SMT	51.3	0.085
		online-A	46.7	-0.109		–	online-A	42.6	-0.252		–	online-A	51.0	0.041
		HY-SMT	45.8	-0.115		$1^{\uparrow}$	HY-AH	40.6	-0.290		$2^{\downarrow}$	AaltoHnmtFlatcat	49.8	0.031
		HY-AH	43.5	-0.192		$1^{\downarrow}$	HY-SMT	40.4	-0.310		$1^{\uparrow}$	jhu-pbmt	46.8	-0.078
		jhu-pbmt	43.4	-0.204		–	jhu-pbmt	40.5	-0.312		$1^{\downarrow}$	HY-AH	47.0	-0.078
	11	TALP-UPC	40.8	-0.298	11	–	TALP-UPC	37.5	-0.413	11	–	TALP-UPC	44.1	-0.183
	12	apertium	8.0	-1.428	12	–	apertium	11.8	-1.293	12	–	apertium	4.4	-1.554
wmt18	1	NICT	64.7	0.521	1	–	NICT	57.0	0.251	1	–	NICT	72.7	0.800
		HY-NMT-en-fi	63.1	0.466		–	HY-NMT-en-fi	56.5	0.232		–	HY-NMT-en-fi	69.6	0.696
	3	uedin	59.2	0.324	3	$1^{\uparrow}$	Aalto	52.8	0.073		–	uedin	67.5	0.636
		Aalto	58.3	0.271		$1^{\uparrow}$	HY-NMTtwostep	52.3	0.045	4	$1^{\uparrow}$	HY-NMTtwostep	64.0	0.492
		HY-NMTtwostep	57.9	0.258		$1^{\uparrow}$	talp-upc	52.4	0.044		$1^{\downarrow}$	Aalto	64.0	0.477
		talp-upc	57.4	0.238		$3^{\downarrow}$	uedin	51.6	0.033		–	talp-upc	62.3	0.430
		CUNI-Kocmi	55.9	0.184		–	CUNI-Kocmi	51.5	0.016		$1^{\uparrow}$	online-B	63.0	0.421
		online-B	56.6	0.183		–	online-B	51.0	-0.027		$1^{\downarrow}$	CUNI-Kocmi	60.8	0.368
	9	online-A	45.9	-0.212	9	$1^{\uparrow}$	online-G	41.2	-0.375	9	–	online-A	52.7	0.032
		online-G	45.3	-0.233	10	$1^{\downarrow}$	online-A	39.6	-0.438	10	–	online-G	50.1	-0.070
	11	HY-SMT-en-fi	42.7	-0.334		$1^{\uparrow}$	HY-AH-en-fi	38.9	-0.460		–	HY-SMT-en-fi	48.7	-0.123
		HY-AH-en-fi	41.5	-0.369		$1^{\downarrow}$	HY-SMT-en-fi	37.1	-0.528	12	–	HY-AH-en-fi	44.3	-0.272

Table 7: Results of the English→Finnish language direction with WMT, ORG, and TRS. Systems are ordered by standardized mean DA score. If a system does not contain a rank, it indicates that it shares the same cluster as the system above it. Clusters are obtained according to Wilcoxon rank-sum test at p-level $p \leq 0.05$. Indicated in the [↑↓] column are the changes in absolute ranking (i.e. how many positions it goes up or down).

German→English

	#	SYSTEM	RAW.WMT	Z.WMT	#	↑↓	SYSTEM	RAW.ORG	Z.ORG	#	↑↓	SYSTEM	RAW.TRS	Z.TRS
wmt16	1	uedin-nmt	75.8	0.204	1	–	uedin-nmt	71.7	0.051	1	–	uedin-nmt	79.9	0.357
	2	online-A	72.7	0.095	2	–	online-A	68.4	-0.046	2	$1^{\uparrow}$	online-B	77.2	0.251
		online-B	72.2	0.086		–	online-B	67.3	-0.079		$1^{\uparrow}$	uedin-syntax	76.4	0.240
		uedin-syntax	71.5	0.065		–	uedin-syntax	66.6	-0.108		$1^{\uparrow}$	KIT	76.6	0.237
		KIT	71.4	0.062		–	KIT	66.2	-0.112		$3^{\downarrow}$	online-A	77.0	0.235
		uedin-pbmt	70.9	0.042		$2^{\uparrow}$	online-G	66.1	-0.120		–	uedin-pbmt	75.6	0.204
		jhu-pbmt	70.5	0.019		$1^{\downarrow}$	uedin-pbmt	66.1	-0.122		–	jhu-pbmt	74.6	0.171
		online-G	70.2	0.009		$1^{\downarrow}$	jhu-pbmt	66.3	-0.133		–	online-G	74.2	0.139
	9	online-F	64.0	-0.204	9	–	online-F	61.4	-0.291	9	–	online-F	66.6	-0.118
		jhu-syntax	62.4	-0.261	10	–	jhu-syntax	58.4	-0.395		–	jhu-syntax	66.4	-0.127
wmt17	1	online-B	78.2	0.213	1	–	online-B	75.8	0.125	1	–	online-B	80.4	0.298
	2	online-A	76.6	0.169	2	$1^{\uparrow}$	KIT	73.7	0.071		$2^{\uparrow}$	uedin-nmt	80.4	0.294
		KIT	76.6	0.165		$1^{\downarrow}$	online-A	74.1	0.069		$1^{\downarrow}$	online-A	79.1	0.269
		uedin-nmt	76.6	0.162		–	uedin-nmt	72.8	0.029		$1^{\downarrow}$	KIT	79.6	0.262
		RWTH-nmt	75.8	0.131		–	RWTH-nmt	73.0	0.021		–	RWTH-nmt	78.7	0.240
		SYSTRAN	74.5	0.098		–	SYSTRAN	71.0	-0.021		–	SYSTRAN	78.0	0.220
	7	LIUM-NMT	72.9	0.029		–	LIUM-NMT	70.4	-0.050	7	–	LIUM-NMT	75.5	0.110
	8	TALP-UPC	70.2	-0.058	8	–	TALP-UPC	67.0	-0.162	8	$1^{\uparrow}$	online-G	74.0	0.080
		online-G	69.8	-0.072		$1^{\uparrow}$	C-3MA	66.3	-0.210		$1^{\downarrow}$	TALP-UPC	73.3	0.044
		C-3MA	68.6	-0.103		$1^{\downarrow}$	online-G	65.6	-0.227		–	C-3MA	70.9	0.004
	11	online-F	64.1	-0.260	11	–	online-F	62.5	-0.325	11	–	online-F	65.9	-0.192
wmt18	1	RWTH	79.9	0.413	1	–	RWTH	76.1	0.281	1	$1^{\uparrow}$	UCAM	84.2	0.553
		UCAM	79.4	0.395	2	–	UCAM	74.6	0.234		$1^{\downarrow}$	RWTH	83.5	0.540
		NTT	78.2	0.359		$1^{\uparrow}$	online-B	72.8	0.199		–	NTT	83.0	0.523
		online-B	77.3	0.346		$1^{\downarrow}$	NTT	73.4	0.196		$2^{\uparrow}$	JHU	82.7	0.504
		MLLP-UPV	77.4	0.321		$3^{\uparrow}$	online-Y	72.8	0.181		$1^{\downarrow}$	online-B	81.8	0.497
		JHU	77.0	0.317		$1^{\downarrow}$	MLLP-UPV	73.4	0.179		$4^{\uparrow}$	uedin	82.2	0.494
		Ubiqus-NMT	76.9	0.315		–	Ubiqus-NMT	72.4	0.172		$2^{\downarrow}$	MLLP-UPV	81.6	0.471
		online-Y	76.7	0.310		$1^{\uparrow}$	online-A	71.4	0.126		$1^{\downarrow}$	Ubiqus-NMT	81.5	0.458
	9	online-A	75.7	0.268		$3^{\downarrow}$	JHU	71.0	0.120		$1^{\downarrow}$	online-Y	80.6	0.440
		uedin	75.4	0.261	10	–	uedin	68.7	0.032		$1^{\downarrow}$	online-A	80.1	0.411
	11	LMU-nmt	72.5	0.162		$1^{\uparrow}$	NJUNMT-private	68.9	0.029		–	LMU-nmt	78.7	0.364
		NJUNMT-private	72.2	0.149	12	$1^{\downarrow}$	LMU-nmt	66.3	-0.035	12	–	NJUNMT-private	75.6	0.270
	13	online-G	65.2	-0.074	13	–	online-G	59.8	-0.244	13	–	online-G	70.4	0.092
	14	online-F	58.5	-0.296	14	–	online-F	56.1	-0.378	14	–	online-F	60.8	-0.214
	15	RWTH-UNSUPER	45.4	-0.752	15	–	RWTH-UNSUPER	41.1	-0.883	15	–	RWTH-UNSUPER	49.6	-0.624
	16	LMU-unsup	42.7	-0.835	16	–	LMU-unsup	38.7	-0.972	16	–	LMU-unsup	46.7	-0.697

Table 8: Results of the German→English language direction with WMT, ORG, and TRS. Systems are ordered by standardized mean DA score. If a system does not contain a rank, it indicates that it shares the same cluster as the system above it. Clusters are obtained according to Wilcoxon rank-sum test at p-level $p \leq 0.05$. Indicated in the [↑↓] column are the changes in absolute ranking (i.e. how many positions it goes up or down).

English→German

	#	SYSTEM	RAW.WMT	Z.WMT	#	↑↓	SYSTEM	RAW.ORG	Z.ORG	#	↑↓	SYSTEM	RAW.TRS	Z.TRS
wmt17	1	LMU-nmt-reranked	72.9	0.257	1	–	LMU-nmt-reranked	68.8	0.101	1	–	LMU-nmt-reranked	77.3	0.423
	2	online-B	70.2	0.158		–	online-B	66.9	0.052		$1^{\uparrow}$	uedin-nmt	75.9	0.356
		uedin-nmt	69.8	0.139		–	uedin-nmt	65.0	-0.036		$1^{\downarrow}$	online-B	74.4	0.294
		SYSTRAN	68.9	0.092		–	SYSTRAN	64.4	-0.059		$1^{\uparrow}$	LMU-nmt-single	73.3	0.280
		LMU-nmt-single	66.9	0.035		$4^{\uparrow}$	RWTH-nmt	61.9	-0.149		$1^{\downarrow}$	SYSTRAN	73.9	0.256
		KIT	66.7	0.022		$1^{\uparrow}$	xmu	62.1	-0.151		–	KIT	72.2	0.238
		xmu	66.4	0.015		$2^{\downarrow}$	LMU-nmt-single	61.7	-0.164		$1^{\uparrow}$	LIUM-NMT	73.0	0.238
		LIUM-NMT	66.6	0.006		–	LIUM-NMT	61.7	-0.172		$1^{\downarrow}$	xmu	70.3	0.165
		RWTH-nmt	66.0	-0.003		$3^{\downarrow}$	KIT	61.7	-0.174		–	RWTH-nmt	70.6	0.162
	10	online-A	60.1	-0.233	10	$1^{\uparrow}$	PROMT-Rule-based	55.6	-0.406	10	–	online-A	65.2	-0.041
		PROMT-Rule-based	60.3	-0.234		$2^{\uparrow}$	fbk-nmt-combi	55.5	-0.406		–	PROMT-Rule-based	64.9	-0.064
		C-3MA	58.9	-0.270		$2^{\downarrow}$	online-A	55.2	-0.418		–	C-3MA	63.8	-0.082
		fbk-nmt-combi	58.1	-0.301		$1^{\downarrow}$	C-3MA	54.6	-0.437		–	fbk-nmt-combi	61.5	-0.162
		TALP-UPC	55.2	-0.391	14	$1^{\uparrow}$	online-F	51.5	-0.570		–	TALP-UPC	60.5	-0.184
		online-F	54.9	-0.440		$1^{\downarrow}$	TALP-UPC	50.3	-0.585		–	online-F	58.5	-0.303
		online-G	53.2	-0.491		–	online-G	48.8	-0.660		–	online-G	57.3	-0.332
wmt18	1	online-Z	85.5	0.653	1	–	online-Z	83.6	0.587	1	–	online-Z	87.4	0.719
	2	online-B	82.2	0.561		–	online-B	81.9	0.544	2	$3^{\uparrow}$	UCAM	84.1	0.599
		Microsoft-Marian	81.9	0.551		–	Microsoft-Marian	81.6	0.533		$1^{\downarrow}$	online-B	82.5	0.578
		MMT-production	81.6	0.539		–	MMT-production	81.5	0.522		$2^{\uparrow}$	NTT	82.5	0.578
		UCAM	82.3	0.537		–	UCAM	80.3	0.475		$2^{\downarrow}$	Microsoft-Marian	82.2	0.568
		NTT	80.2	0.491		–	NTT	78.1	0.409		$2^{\downarrow}$	MMT-production	81.7	0.556
		KIT	79.3	0.454		$1^{\uparrow}$	online-Y	77.7	0.403		–	KIT	81.1	0.525
	8	online-Y	77.7	0.396		$1^{\downarrow}$	KIT	77.5	0.383		$1^{\uparrow}$	JHU	80.2	0.497
		JHU	76.7	0.377		$1^{\uparrow}$	uedin	74.3	0.298	9	$1^{\uparrow}$	uedin	78.3	0.405
		uedin	76.3	0.352		$1^{\downarrow}$	JHU	73.5	0.265		$2^{\downarrow}$	online-Y	77.7	0.389
	11	LMU-nmt	71.8	0.213	11	–	LMU-nmt	68.7	0.103		–	LMU-nmt	74.6	0.317
	12	online-A	67.4	0.060	12	–	online-A	62.9	-0.087	12	–	online-A	71.9	0.208
	13	online-F	53.2	-0.385	13	–	online-F	50.7	-0.463	13	–	online-F	55.6	-0.309
		online-G	53.8	-0.416		–	online-G	51.3	-0.505		–	online-G	56.4	-0.326
	15	RWTH-UNSUPER	36.7	-0.966	15	–	RWTH-UNSUPER	34.8	-1.002	15	–	RWTH-UNSUPER	38.7	-0.930
	16	LMU-unsup	32.6	-1.122	16	–	LMU-unsup	30.0	-1.193	16	–	LMU-unsup	35.1	-1.056

Table 9: Results of the English→German language direction with WMT, ORG, and TRS. Systems are ordered by standardized mean DA score. If a system does not contain a rank, it indicates that it shares the same cluster as the system above it. Clusters are obtained according to Wilcoxon rank-sum test at p-level $p \leq 0.05$. Indicated in the [↑↓] column are the changes in absolute ranking (i.e. how many positions it goes up or down).

Latvian→English

	#	SYSTEM	RAW.WMT	Z.WMT	#	↑↓	SYSTEM	RAW.ORG	Z.ORG	#	↑↓	SYSTEM	RAW.TRS	Z.TRS
wmt17	1	online-B	76.2	0.266	1	–	online-B	75.8	0.266	1	$1^{\uparrow}$	tilde-nc-nmt-smt	76.8	0.268
		tilde-nc-nmt-smt	76.2	0.245		–	tilde-nc-nmt-smt	75.5	0.222		$1^{\downarrow}$	online-B	76.5	0.267
	3	uedin-nmt	71.4	0.087	3	$1^{\uparrow}$	tilde-c-nmt-smt	69.8	0.043	3	–	uedin-nmt	74.0	0.168
		tilde-c-nmt-smt	71.0	0.083		$1^{\downarrow}$	uedin-nmt	68.8	0.007	4	–	tilde-c-nmt-smt	72.1	0.121
	5	online-A	67.3	-0.039	5	–	online-A	64.5	-0.142		–	online-A	70.0	0.062
	6	jhu-pbmt	64.4	-0.137		–	jhu-pbmt	63.1	-0.185	6	–	jhu-pbmt	65.8	-0.089
	7	C-3MA	63.4	-0.187		–	C-3MA	62.5	-0.223		$1^{\uparrow}$	Hunter-MT	63.9	-0.134
		Hunter-MT	62.2	-0.199		–	Hunter-MT	60.3	-0.264		$1^{\downarrow}$	C-3MA	64.2	-0.153
	9	PJATK	56.3	-0.436	9	–	PJATK	53.5	-0.554	9	–	PJATK	59.1	-0.316

Table 10: Results of the Latvian→English language direction with WMT, ORG, and TRS. Systems are ordered by standardized mean DA score. If a system does not contain a rank, it indicates that it shares the same cluster as the system above it. Clusters are obtained according to Wilcoxon rank-sum test at p-level $p \leq 0.05$. Indicated in the [↑↓] column are the changes in absolute ranking (i.e. how many positions it goes up or down).

English→Latvian

	#	SYSTEM	RAW.WMT	Z.WMT	#	↑↓	SYSTEM	RAW.ORG	Z.ORG	#	↑↓	SYSTEM	RAW.TRS	Z.TRS
wmt17	1	tilde-nc-nmt-smt	54.4	0.196	1	–	tilde-nc-nmt-smt	43.2	-0.168	1	–	tilde-nc-nmt-smt	66.1	0.579
		online-B	51.6	0.121		–	online-B	40.6	-0.222		$1^{\uparrow}$	tilde-c-nmt-smt	64.9	0.519
		tilde-c-nmt-smt	51.1	0.104		$1^{\uparrow}$	limsi-factored-norm	41.3	-0.235		$1^{\downarrow}$	online-B	63.1	0.484
		limsi-factored-norm	50.8	0.075		$3^{\uparrow}$	usfd-consensus-kit	40.4	-0.244		$1^{\uparrow}$	usfd-consensus-qt21	61.0	0.413
		usfd-consensus-qt21	50.0	0.058		$2^{\downarrow}$	tilde-c-nmt-smt	39.2	-0.255		$1^{\downarrow}$	limsi-factored-norm	61.0	0.410
		QT21-System-Combi	47.1	-0.014		$1^{\downarrow}$	usfd-consensus-qt21	40.0	-0.264		–	QT21-System-Combi	58.8	0.346
		usfd-consensus-kit	47.3	-0.027		$2^{\uparrow}$	uedin-nmt	39.1	-0.271		–	usfd-consensus-kit	54.7	0.205
		KIT	45.7	-0.063		–	KIT	37.1	-0.321		–	KIT	54.5	0.200
		uedin-nmt	45.2	-0.072		$3^{\downarrow}$	QT21-System-Combi	36.8	-0.334		$1^{\uparrow}$	tilde-nc-smt	55.1	0.183
		tilde-nc-smt	44.9	-0.099		–	tilde-nc-smt	34.5	-0.387		$1^{\downarrow}$	uedin-nmt	52.6	0.168
		LIUM-FNMT	43.2	-0.157		$1^{\uparrow}$	LIUM-NMT	35.7	-0.461		–	LIUM-FNMT	51.7	0.125
		LIUM-NMT	43.0	-0.198		$1^{\downarrow}$	LIUM-FNMT	34.0	-0.464		–	LIUM-NMT	50.0	0.055
		HY-HNMT	40.1	-0.253		–	HY-HNMT	30.2	-0.572		–	HY-HNMT	47.8	-0.005
		online-A	37.5	-0.341		–	online-A	30.0	-0.573		$1^{\uparrow}$	jhu-pbmt	44.5	-0.099
		jhu-pbmt	36.1	-0.368		–	jhu-pbmt	28.2	-0.618		$1^{\downarrow}$	online-A	44.5	-0.124
		C-3MA	33.3	-0.457	16	–	C-3MA	24.6	-0.735		–	C-3MA	41.3	-0.201
	17	PJATK	18.8	-0.947	17	–	PJATK	13.9	-1.138	17	–	PJATK	23.8	-0.752

Table 11: Results of the English→Latvian language direction with WMT, ORG, and TRS. Systems are ordered by standardized mean DA score. If a system does not contain a rank, it indicates that it shares the same cluster as the system above it. Clusters are obtained according to Wilcoxon rank-sum test at p-level $p \leq 0.05$. Indicated in the [↑↓] column are the changes in absolute ranking (i.e. how many positions it goes up or down).

Romanian→English

	#	SYSTEM	RAW.WMT	Z.WMT	#	↑↓	SYSTEM	RAW.ORG	Z.ORG	#	↑↓	SYSTEM	RAW.TRS	Z.TRS
wmt16	1	online-B	73.9	0.129	1	–	online-B	73.5	0.117	1	–	online-B	74.4	0.140
	2	uedin-nmt	71.2	0.044	2	–	uedin-nmt	70.9	0.037	2	$1^{\uparrow}$	uedin-pbmt	72.1	0.063
		uedin-pbmt	71.0	0.025		–	uedin-pbmt	69.9	-0.013		$2^{\uparrow}$	online-A	72.2	0.058
		uedin-syntax	69.9	-0.000		–	uedin-syntax	68.6	-0.031		$2^{\downarrow}$	uedin-nmt	71.4	0.052
		online-A	69.7	-0.012		–	online-A	67.2	-0.082		$1^{\downarrow}$	uedin-syntax	71.2	0.030
	6	LIMSI	66.7	-0.123	6	–	LIMSI	63.1	-0.257		–	LIMSI	70.3	0.012
		jhu-pbmt	65.7	-0.160		–	jhu-pbmt	60.6	-0.306		–	jhu-pbmt	70.8	-0.012

Table 12: Results of the Romanian→English language direction with WMT, ORG, and TRS. Systems are ordered by standardized mean DA score. If a system does not contain a rank, it indicates that it shares the same cluster as the system above it. Clusters are obtained according to Wilcoxon rank-sum test at p-level $p \leq 0.05$. Indicated in the [↑↓] column are the changes in absolute ranking (i.e. how many positions it goes up or down).

Russian→English

	#	SYSTEM	RAW.WMT	Z.WMT	#	↑↓	SYSTEM	RAW.ORG	Z.ORG	#	↑↓	SYSTEM	RAW.TRS	Z.TRS
wmt16	1	online-G	74.2	0.115	1	$4^{\uparrow}$	PROMT-Rule-based	73.0	0.072	1	–	online-G	76.0	0.172
		AMU-UEDIN	73.3	0.103		$1^{\downarrow}$	online-G	72.5	0.058		–	AMU-UEDIN	74.6	0.155
		online-B	72.8	0.083		$1^{\downarrow}$	AMU-UEDIN	72.0	0.051		–	online-B	74.8	0.142
		NRC	72.7	0.060		$1^{\downarrow}$	online-B	70.8	0.025		–	NRC	75.0	0.140
	5	PROMT-Rule-based	72.1	0.044		$1^{\downarrow}$	NRC	70.3	-0.020	5	$1^{\uparrow}$	uedin-nmt	72.3	0.061
		uedin-nmt	71.1	0.011		–	uedin-nmt	70.0	-0.039		$1^{\uparrow}$	online-A	72.7	0.055
		online-A	70.8	-0.007		–	online-A	68.9	-0.069		$1^{\uparrow}$	AFRL-MITLL-Phrase	72.2	0.030
		AFRL-MITLL-Phrase	70.1	-0.040		–	AFRL-MITLL-Phrase	67.9	-0.111	8	$3^{\downarrow}$	PROMT-Rule-based	71.3	0.016
		AFRL-MITLL-contrast	69.3	-0.071		–	AFRL-MITLL-contrast	68.2	-0.125		–	AFRL-MITLL-contrast	70.5	-0.018
	10	online-F	61.8	-0.322	10	–	online-F	62.0	-0.295	10	–	online-F	61.6	-0.349
wmt17	1	online-B	82.0	0.271	1	–	online-B	81.3	0.255	1	–	online-B	82.6	0.288
	2	online-G	77.6	0.126	2	–	online-G	76.0	0.052	2	–	online-G	79.1	0.196
	3	NRC	76.5	0.081		–	NRC	74.4	-0.001		–	NRC	78.7	0.161
		online-A	76.1	0.057		–	online-A	74.3	-0.004		–	online-A	78.0	0.118
		afrl-mitll-syscomb	74.9	0.017		$1^{\uparrow}$	afrl-mitll-opennmt	73.8	-0.007		–	afrl-mitll-syscomb	76.8	0.087
		afrl-mitll-opennmt	74.6	0.005		$1^{\downarrow}$	afrl-mitll-syscomb	73.1	-0.053	6	$2^{\uparrow}$	jhu-pbmt	77.1	0.071
		uedin-nmt	74.2	0.002		–	uedin-nmt	72.3	-0.062		–	uedin-nmt	76.1	0.062
		jhu-pbmt	74.7	-0.011		–	jhu-pbmt	72.4	-0.091		$1^{\downarrow}$	afrl-mitll-opennmt	75.3	0.017
	9	online-F	65.9	-0.288	9	–	online-F	65.8	-0.290	9	–	online-F	66.0	-0.287
wmt18	1	Alibaba	81.0	0.215	1	–	Alibaba	80.9	0.197	1	–	Alibaba	81.0	0.232
		online-B	80.3	0.192		–	online-B	80.2	0.185		–	online-B	80.3	0.199
		online-G	79.6	0.170		–	online-G	78.8	0.143		–	online-G	80.3	0.197
	4	uedin	77.5	0.110		–	uedin	76.6	0.080	4	–	uedin	78.3	0.141
	5	online-A	76.2	0.034	5	–	online-A	75.7	0.010	5	–	online-A	76.6	0.058
	6	afrl-ruen-syscomb	74.1	-0.014		$1^{\uparrow}$	JHU	73.6	-0.026		–	afrl-ruen-syscomb	74.4	0.003
		JHU	73.7	-0.027		$1^{\downarrow}$	afrl-ruen-syscomb	73.7	-0.032		–	JHU	73.8	-0.029
	8	online-F	64.2	-0.398	8	–	online-F	66.0	-0.322	8	–	online-F	62.5	-0.475

Table 13: Results of the Russian→English language direction with WMT, ORG, and TRS. Systems are ordered by standardized mean DA score. If a system does not contain a rank, it indicates that it shares the same cluster as the system above it. Clusters are obtained according to Wilcoxon rank-sum test at p-level $p \leq 0.05$. Indicated in the [↑↓] column are the changes in absolute ranking (i.e. how many positions it goes up or down).

English→Russian

	#	SYSTEM	RAW.WMT	Z.WMT	#	↑↓	SYSTEM	RAW.ORG	Z.ORG	#	↑↓	SYSTEM	RAW.TRS	Z.TRS
wmt17	1	online-B	75.4	0.402	1	–	online-B	69.6	0.202	1	–	online-B	81.2	0.601
	2	uedin-nmt	68.2	0.166	2	–	uedin-nmt	60.4	-0.091	2	–	uedin-nmt	76.0	0.424
	3	online-H	66.5	0.105		$1^{\uparrow}$	PROMT-Rule-based	60.4	-0.105		–	online-H	74.5	0.384
	4	PROMT-Rule-based	65.9	0.080		$1^{\uparrow}$	online-A	59.2	-0.137	4	$2^{\uparrow}$	online-G	73.6	0.326
		online-A	65.2	0.061		$2^{\downarrow}$	online-H	58.9	-0.159	5	$1^{\downarrow}$	PROMT-Rule-based	71.6	0.273
		online-G	65.2	0.054	6	–	online-G	56.9	-0.214		$1^{\downarrow}$	online-A	71.1	0.255
	7	jhu-pbmt	62.6	-0.018		–	jhu-pbmt	54.6	-0.273		–	jhu-pbmt	70.6	0.240
	8	afrl-mitll-backtrans	57.3	-0.194	7	–	afrl-mitll-backtrans	50.7	-0.418	8	–	afrl-mitll-backtrans	63.9	0.032
	9	online-F	46.5	-0.568	8	–	online-F	41.5	-0.740	9	–	online-F	51.4	-0.405
wmt18	1	Alibaba-ensemble	72.0	0.352	1	–	Alibaba-ensemble	64.6	0.113	1	–	Alibaba-ensemble	79.4	0.592
		online-G	71.4	0.324		–	online-G	64.3	0.075		–	online-G	78.5	0.570
	3	online-B	66.8	0.159	3	–	online-B	60.1	-0.049	3	$1^{\uparrow}$	uedin	73.2	0.389
		uedin	66.0	0.144		–	uedin	58.8	-0.101		$1^{\downarrow}$	online-B	73.4	0.365
		PROMT-Marian	64.9	0.115		–	PROMT-Marian	58.0	-0.114		–	PROMT-Marian	72.1	0.355
	6	PROMT-OpenNMT	63.9	0.066		–	PROMT-OpenNMT	56.5	-0.155		$1^{\uparrow}$	online-A	70.8	0.292
	7	online-A	62.2	-0.004	7	$1^{\uparrow}$	PROMT-Rule-based	53.7	-0.242		$1^{\downarrow}$	PROMT-OpenNMT	71.0	0.279
	8	PROMT-Rule-based	59.1	-0.075		$1^{\downarrow}$	online-A	53.8	-0.292	8	–	PROMT-Rule-based	64.8	0.097
	9	online-F	44.5	-0.580	9	–	online-F	42.5	-0.656	9	–	online-F	46.5	-0.502

Table 14: Results of the English→Russian language direction with WMT, ORG, and TRS. Systems are ordered by standardized mean DA score. If a system does not contain a rank, it indicates that it shares the same cluster as the system above it. Clusters are obtained according to Wilcoxon rank-sum test at p-level $p \leq 0.05$. Indicated in the [↑↓] column are the changes in absolute ranking (i.e. how many positions it goes up or down).

Turkish→English

	#	SYSTEM	RAW.WMT	Z.WMT	#	↑↓	SYSTEM	RAW.ORG	Z.ORG	#	↑↓	SYSTEM	RAW.TRS	Z.TRS
wmt16	1	online-B	57.1	0.163	1	–	online-B	55.5	0.120	1	–	online-B	58.7	0.205
	2	online-G	55.0	0.109	2	–	online-G	53.4	0.057		–	online-G	56.4	0.157
	3	online-A	52.2	0.002		–	online-A	51.5	-0.009	3	–	online-A	52.9	0.012
	4	tbtk-syscomb	49.6	-0.077	4	$2^{\uparrow}$	dvorkanton	48.8	-0.120		$1^{\uparrow}$	PROMT-SMT	51.5	-0.015
		PROMT-SMT	49.2	-0.079		$1^{\downarrow}$	tbtk-syscomb	48.8	-0.140		$1^{\downarrow}$	tbtk-syscomb	50.3	-0.017
		dvorkanton	49.5	-0.088		$1^{\downarrow}$	PROMT-SMT	46.9	-0.144		–	dvorkanton	50.2	-0.057
	7	jhu-pbmt	41.0	-0.355	7	–	jhu-pbmt	40.4	-0.381	7	$1^{\uparrow}$	jhu-syntax	41.9	-0.303
		jhu-syntax	40.8	-0.364		$1^{\uparrow}$	ParFDA	39.7	-0.390		$1^{\downarrow}$	jhu-pbmt	41.5	-0.329
		ParFDA	40.5	-0.367		$1^{\downarrow}$	jhu-syntax	39.8	-0.422		–	ParFDA	41.2	-0.345
wmt17	1	online-B	68.8	0.294	1	–	online-B	65.0	0.171	1	–	online-B	72.7	0.417
		online-A	68.5	0.282		–	online-A	64.5	0.153		–	online-A	72.5	0.407
	3	uedin-nmt	61.1	0.050	3	–	uedin-nmt	57.8	-0.051	3	–	uedin-nmt	64.3	0.148
	4	online-G	58.6	-0.029		–	online-G	57.1	-0.094	4	$1^{\uparrow}$	afrl-mitll-m2w-nr1	61.4	0.057
		afrl-mitll-m2w-nr1	58.0	-0.083	5	$2^{\uparrow}$	LIUM-NMT	54.6	-0.168		$1^{\uparrow}$	afrl-mitll-syscomb	60.3	0.040
		afrl-mitll-syscomb	57.0	-0.093		$1^{\downarrow}$	afrl-mitll-m2w-nr1	54.6	-0.220		$2^{\downarrow}$	online-G	60.1	0.036
		LIUM-NMT	56.7	-0.097		$1^{\downarrow}$	afrl-mitll-syscomb	53.7	-0.224		–	LIUM-NMT	58.8	-0.028
	8	PROMT-SMT	53.5	-0.183	8	–	PROMT-SMT	52.6	-0.227	8	–	PROMT-SMT	54.5	-0.139
	9	jhu-pbmt	46.4	-0.436	9	–	jhu-pbmt	44.9	-0.463	9	–	jhu-pbmt	48.0	-0.408
		JAIST	45.5	-0.475		–	JAIST	45.1	-0.494	10	–	JAIST	46.0	-0.456
wmt18	1	online-G	74.3	0.045	1	–	online-G	71.1	-0.084	1	$1^{\uparrow}$	online-A	78.2	0.192
		online-A	74.3	0.040		$1^{\uparrow}$	online-B	70.6	-0.112		$1^{\downarrow}$	online-G	77.4	0.174
		online-B	73.0	-0.004		$1^{\downarrow}$	online-A	70.2	-0.115	3	–	online-B	75.3	0.100
		uedin	71.7	-0.053		–	uedin	69.0	-0.175		$1^{\uparrow}$	NICT	74.5	0.081
		NICT	71.6	-0.055		–	NICT	68.7	-0.192		$1^{\downarrow}$	uedin	74.4	0.067

Table 15: Results of the Turkish→English language direction with WMT, ORG, and TRS. Systems are ordered by standardized mean DA score. If a system does not contain a rank, it indicates that it shares the same cluster as the system above it. Clusters are obtained according to Wilcoxon rank-sum test at p-level $p \leq 0.05$. Indicated in the [↑↓] column are the changes in absolute ranking (i.e. how many positions it goes up or down).

English→Turkish

	#	SYSTEM	RAW.WMT	Z.WMT	#	↑↓	SYSTEM	RAW.ORG	Z.ORG	#	↑↓	SYSTEM	RAW.TRS	Z.TRS
wmt17	1	online-B	53.4	0.513	1	–	online-B	40.0	0.131	1	–	online-B	65.2	0.848
	2	uedin-nmt	44.0	0.206	2	–	uedin-nmt	32.5	-0.134	2	–	uedin-nmt	56.5	0.576
	3	online-A	39.1	0.071		–	online-A	28.6	-0.241		–	online-A	50.3	0.406
		online-G	35.5	-0.032		–	online-G	27.5	-0.285	4	–	online-G	42.9	0.200
	5	LIUM-NMT	32.2	-0.129		–	LIUM-NMT	23.6	-0.376		–	LIUM-NMT	41.3	0.132
	6	jhu-nmt-lattice	18.0	-0.554	6	–	jhu-nmt-lattice	14.3	-0.654	6	–	jhu-nmt-lattice	21.2	-0.469
		jhu-pbmt	16.7	-0.597		$1^{\uparrow}$	JAIST	12.4	-0.690		–	jhu-pbmt	20.5	-0.484
		JAIST	15.7	-0.602		$1^{\downarrow}$	jhu-pbmt	12.6	-0.717		–	JAIST	19.3	-0.504
wmt18	1	online-B	66.3	0.277	1	$1^{\uparrow}$	uedin	62.2	0.149	1	–	online-B	71.8	0.444
		uedin	63.6	0.222		$1^{\downarrow}$	online-B	61.1	0.117		$1^{\uparrow}$	alibaba-ensemble-A	67.9	0.348
		alibaba-ensemble-A	63.5	0.216		–	alibaba-ensemble-A	59.6	0.097		$1^{\downarrow}$	uedin	65.2	0.304
		NICT	62.0	0.128		–	NICT	59.5	0.037		$1^{\uparrow}$	alibaba-ensemble-B	65.3	0.270
		alibaba-ensemble-B	60.1	0.111		–	alibaba-ensemble-B	55.5	-0.030		$1^{\uparrow}$	online-G	65.5	0.264
		online-G	60.1	0.058		–	online-G	54.7	-0.145		$2^{\downarrow}$	NICT	64.8	0.229
	7	RWTH	55.0	-0.060		–	RWTH	52.9	-0.150	7	–	RWTH	57.1	0.029
	8	online-A	49.6	-0.254	8	–	online-A	47.4	-0.331	8	–	online-A	51.9	-0.169

Table 16: Results of the English→Turkish language direction with WMT, ORG, and TRS. Systems are ordered by standardized mean DA score. If a system does not contain a rank, it indicates that it shares the same cluster as the system above it. Clusters are obtained according to Wilcoxon rank-sum test at p-level $p \leq 0.05$. Indicated in the [↑↓] column are the changes in absolute ranking (i.e. how many positions it goes up or down).