A Human Annotation Details

In fig. 6 we provide a screenshot of a website we built for human annotation. You can see there an example of an instance to annotate as well as the questions which were shown to the annotators.

Payment details. Prolific only allows to set a fixed rate for the task completed by the annotator. So, we first estimated how long it takes to annotate 100 instances and to read all the instruction, and we would set the rates accordingly to get an average of 10.5€ per hour. However, in reality some annotators would be faster than anticipated, so the actual average (total payment divided by total time spent) is about 10.7€ per hour.

B Full Automatic Evaluation Results

In table 5 we provide all the automatic evaluation results of all the models on all of the test sets for Russian. We can see that for Russian models all metrics favour the Interno model. In table 6 we provide all the automatic evaluation results of all the models on all of the test sets for English. Only for KELM-E+P we see that different metrics favour different models, however the difference for TER and BERT metrics is not significant.

As according to table 2 KELM-E+P is also balanced regarding the input graph size, we decided to also additionally provide BLEU scores by each graph size for this benchmark for both English and Russian. The graphs in fig. 7 illustrate those results. We see a well expected drop of performance for all English models with a graph size increase. However, we do not see the same drop for Russian models. This may be due to the overall poorer performance of Russian models on unseen data.

C Graph sizes in Kelm-E+P

We provide graph sizes distribution in Kelm-E+P depending on each ratio range in table 7.

D Error Annotation Examples

We show some of the annotated examples of generated texts in English (table 8) and Russian (tables 9 and 10). We highlight there the specific issues which led to annotation of the specific error. Note, that normally each generated text may have several problems, in the table we only highlight the one which illustrates the specific error. We also provide translations for Russian examples (except for "garbage" category).

E Full Qualitative Analysis Results

We present all the ratios of errors and good instances across all models, benchmarks and languages in table 11. We also show a progression of a ratio of each error for each model on KELM-E+P benchmark for English (fig. 8) and Russian (fig. 9).

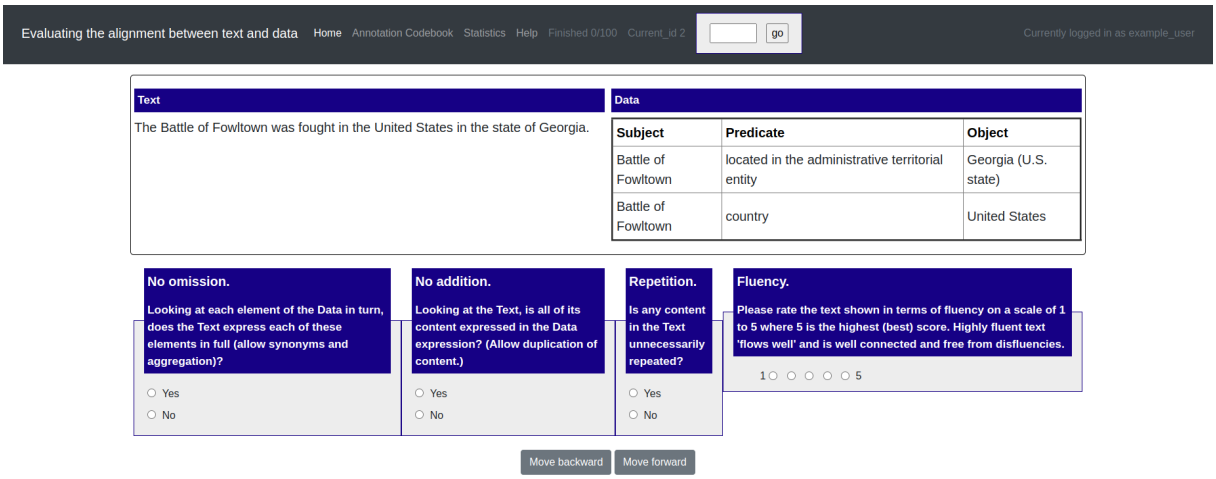

Figure 6: A screenshot from the annotation website we built. Evaluation of a text-data pair.

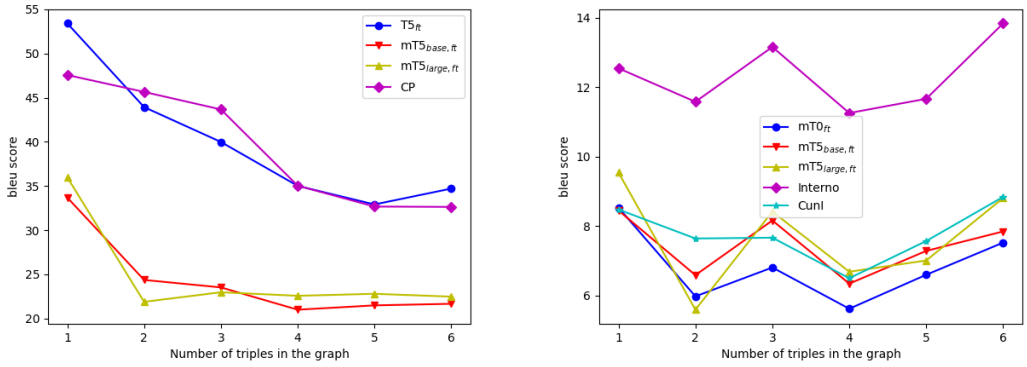

Figure 7: BLEU scores for different graph sizes in KELM-E+Pfor English (left) and Russian (right)

Model	BLEU ↑	chrF++ ↑	TER ↓	BERT P ↑	BERT R ↑	BERT F1 ↑
ru-WebNLG All						
$mT0_{ft}$	52.227	0.685	0.397	0.915	0.910	0.912
$mT5_{base,ft}$	51.861	0.684	0.393	0.916	0.909	0.911
$mT5_{large,ft}$	51.954	0.686	0.401	0.914	0.908	0.910
Interno	**53.668**	**0.694**	**0.373**	**0.921**	**0.913**	**0.916**
CunI	48.503	0.673	0.439	0.902	0.897	0.898
ru-WebNLG E						
$mT0_{ft}$	17.51	0.457	0.777	0.82	0.821	0.82
$mT5_{base,ft}$	17.212	0.438	0.787	0.814	0.81	0.811
$mT5_{large,ft}$	15.1	0.439	0.87	0.792	0.805	0.798
Interno	**22.427**	**0.486**	**0.649**	**0.858**	**0.851**	**0.854**
CunI	17.016	0.426	0.765	0.804	0.805	0.804
ru-WebNLG C						
$mT0_{ft}$	11.193	0.175	0.931	0.754	0.755	0.753
$mT5_{base,ft}$	11.272	0.177	0.854	0.756	0.755	0.754
$mT5_{large,ft}$	11.289	0.176	0.928	0.746	0.748	0.745
Interno	**23.019**	**0.422**	**0.653**	**0.84**	**0.819**	**0.828**
CunI	11.658	0.289	0.814	0.757	0.758	0.757
KELM-E						
$mT0_{ft}$	11.318	0.311	0.876	0.781	**0.79**	0.785
$mT5_{base,ft}$	10.121	0.296	0.943	0.767	0.781	0.773
$mT5_{large,ft}$	10.989	0.306	0.957	0.761	0.776	0.768
Interno	**17.569**	**0.355**	**0.692**	**0.818**	**0.79**	**0.803**
CunI	10.779	0.307	0.893	0.773	0.788	0.78
KELM-E+P						
$mT0_{ft}$	6.811	0.284	1.314	0.73	0.747	0.738
$mT5_{base,ft}$	7.483	0.284	1.178	0.734	0.749	0.741
$mT5_{large,ft}$	7.768	0.284	1.225	0.723	0.742	0.732
Interno	**12.391**	**0.355**	**1.055**	**0.78**	**0.777**	**0.778**
CunI	7.725	0.294	1.08	0.74	0.756	0.747

Table 5: **Automatic evaluation results on Russian benchmarks**

Model	BLEU ↑	chrF++ ↑	TER ↓	BERT P ↑	BERT R ↑	BERT F1 ↑
en-WebNLG						
$T5_{ft}$	52.569	0.680	0.411	0.958	0.955	0.956
$mT5_{base,ft}$	44.163	0.627	0.575	0.942	0.941	0.941
$mT5_{large,ft}$	44.019	0.634	0.558	0.942	0.942	0.941
CP	**53.81**	**0.692**	**0.401**	**0.959**	**0.956**	**0.957**
KELM-E						
$T5_{ft}$	**47.554**	**0.764**	**0.346**	**0.96**	**0.967**	**0.963**
$mT5_{base,ft}$	31.694	0.66	0.645	0.938	0.946	0.942
$mT5_{large,ft}$	32.086	0.675	0.541	0.941	0.947	0.944
CP	46.629	0.755	0.368	0.959	0.965	0.962
KELM-E+P						
$T5_{ft}$	**39.037**	**0.685**	0.566	**0.946**	0.955	0.95
$mT5_{base,ft}$	23.794	0.566	1.007	0.913	0.927	0.92
$mT5_{large,ft}$	24.419	0.581	0.928	0.915	0.931	0.923
CP	38.863	0.683	**0.553**	**0.946**	**0.956**	**0.951**

Table 6: **Automatic evaluation results on English benchmarks**, i.e., WebNLG data (WebNLG: 2020 challenge test set), and two new created benchmarks for text generation from KELM.

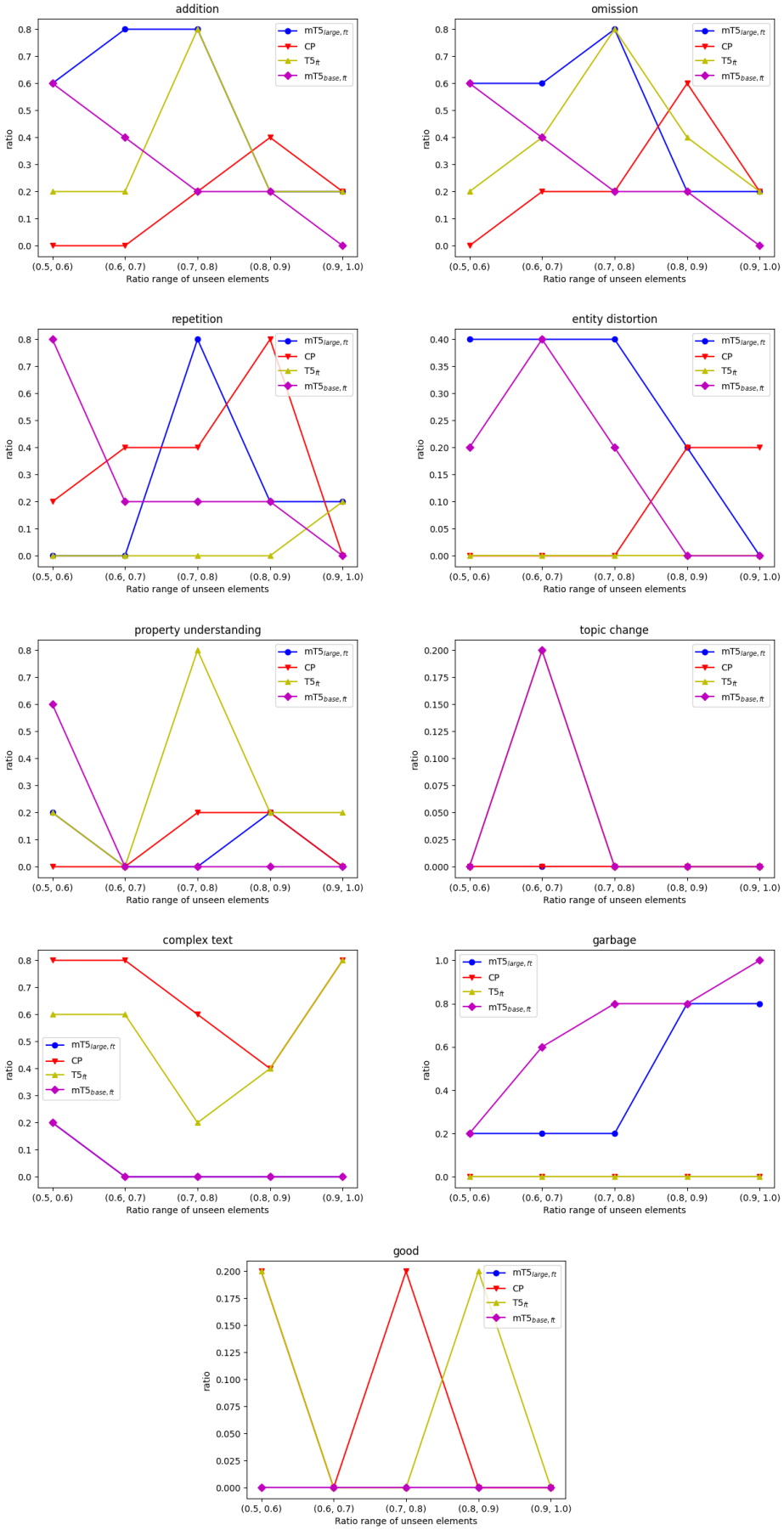

Figure 8: Each error progression on KELM-E+P dataset for English.

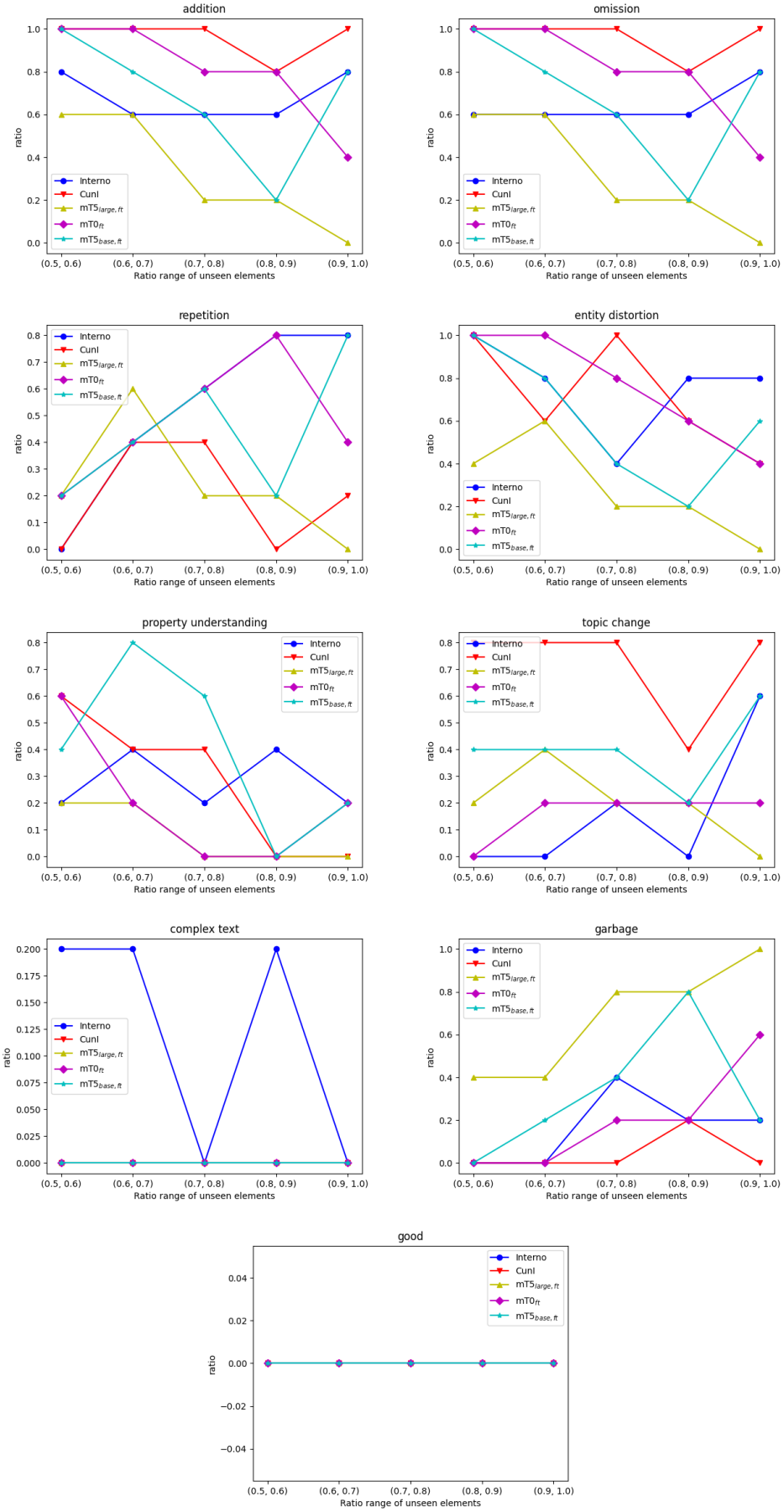

Figure 9: Each error progression on KELM-E+P dataset for Russian.

Ratio \ Graph size	1	2	3	4	5	6
(0.5, 0.6)	0	0	86	33	21	6
(0.6, 0.7)	1	47	91	29	22	21
(0.7, 0.8)	0	0	42	82	109	95
(0.8, 0.9)	0	42	42	49	49	83
(0.9, 1.0)	175	38	34	64	37	13

Table 7: Graph sizes distribution in each ratio range in KELM-E+P

Error	Graph	Generated text
Addition	(*Drenova (Prijepolje)*, *instance of*, *Human settlement*), (*Drenova (Prijepolje)*, *country*, *Serbia*), (*Drenova (Prijepolje)*, *located in area*, *Prijepolje*)	The city of Prijepoli in Serbia is located in the city of Prijepoli. It is a settlement located in the area of Prijepoli.
Omission	(*Rybakowo, Warmian-Masurian Voivodeship, instance of, village of Poland*), (*Rybakowo*, *Warmian-Masurian Voivodeship*, *country*, *Poland*), (*Rybakowo*, *Warmian-Masurian Voivodeship*, *located in area*, *Gmina Srokowo*)	Rybakowo, Warmian-Masurian Voivodeship is located in Gmina Srokowo, Poland.
Repetition	(*Travis Heights, Austin, Texas*, *instance of*, *Neighbourhood*), (*Travis Heights, Austin, Texas*, *country*, *United States*), (*Travis Heights, Austin, Texas*, *located in area*, *Austin, Texas*)	Travis Heights, Austin, Texas is a neighborhood in the U.S.A. and is located in Austin, Texas.
Entity distortion	(*Sandy Lloyd*, *occupation*, *Umpire (cricket)*), (*Sandy Lloyd*, *sport*, *Cricket*), (*Sandy Lloyd*, *occupation*, *Cricketer*), (*Sandy Lloyd*, *sex*, *Male gender*), (*Sandy Lloyd*, *given name*, *Sandy (given name)*), (*Sandy Lloyd*, *family name*, *Lloyd (surname)*)	Sandy Elliot is a Male who was given the nickname 'Sandy'...
Property understanding	(*Elliott, Illinois*, *instance of*, *Village (United States)*), (*Elliott, Illinois*, *country*, *United States*), (*Elliott, Illinois*, *located in area*, *Illinois*)	Elliott is located in Illinois in the United States. It is part of the Village in the United States.
Topic change	(*James Baines (clipper)*, *instance of*, *Ship*), (*James Baines (clipper)*, *country*, *United Kingdom*), (*James Baines (clipper)*, *location of final assembly*, *Boston*)	The clipper James Baines, who was a member of the shipyard of the United Kingdom, was assembled in Boston.
Complex text	(*Palaistra, Florina*, *instance of*, *Human settlement*), (*Palaistra, Florina*, *country*, *Greece*), (*Palaistra, Florina*, *located in area*, *Florina Municipality*)	Palaistra, Florina is located in Florina Municipality, Greece and is an example of human settlement.
Garbage	(*Henuttawy (princess)*, *sibling*, *Menkheperre*), (*Henuttawy (princess)*, *family*, *Twenty-first Dynasty of Egypt*), (*Henuttawy (princess)*, *sibling*, *Masaharta*)	The twelid of the twelid of the twelid of the twelid of the twelid of the twelid of the twelid of the twelid...
Good	(*Brian Mallette*, *weight*, *+84 kilogram*)	Brian Mallette's weight is +84 kilogram.

Table 8: Error annotation examples for English

Error	Graph	Generated text
Addition	(*Leipzig-Schönefeld*, *instance of*, *Ortsteil*), (*Leipzig-Schönefeld*, *country*, *Germany*), (*Leipzig-Schönefeld*, *located in area*, *Leipzig*)	Германский город Лейпциг-Шонефельд находится в районе Лейпцига и находится в области Лейпцига. В городе есть Ортстеил. *Translation:* The German city of Leipzig-Schonefeld is located in the district of Leipzig and is in the region of Leipzig. The city has an Ortsteil.
Omission	(*Hypermarcas*, *industry*, *Pharmaceuticals*), (*Hypermarcas*, *type*, *S.A. (corporation)*), (*Hypermarcas*, *subsidiary*, *Mantecorp*), (*Hypermarcas*, *net income*, *108600000*), (*Hypermarcas*, *number of employees*, *10252*), (*Hypermarcas*, *product*, *Healthcare*)	Компания "Hypermarcas фармацевтическая компания, которая имеет дочернюю компанию "Mantecorp". "Hypermarcas"имеет 10252 сотрудника и годовой доход 108600000. *Translation:* Hypermarcas is a pharmaceutical company which has a subsidiary Mantecorp. Hypermarcas has 10252 employees and has an annual revenue of 108600000.
Repetition	(*Parliamentary Entitlements Act 1990*, *country*, *Australia*), (*Parliamentary Entitlements Act 1990*, *instance of*, *Act of the Parliament of Australia*), (*Parliamentary Entitlements Act 1990*, *applies to jurisdiction*, *Australia*)	Парламент Австралии принял Парламентский акт 1990, который относится к юриспруденции в Австралии. Он был принят Парламентом Австралии и был опубликован в Акте Парламента Австралии. *Translation:* The Parliament of Australia passed the Parliament Act 1990, which relates to jurisprudence in Australia. It was passed by the Parliament of Australia and was published in an Act of the Parliament of Australia.
Entity distortion	(*Bedford Aerodrome*, *runway surface type*, *Concrete*)	Тип поверхности взлетно-посадочной полосы в аэропорту "Берддорф" - бетон. *Translation:* The surface type of the runway at Berddorf Airport is concrete.

Table 9: Error annotation examples for Russian: addition, omission, repetition and entity distortion

Error	**Graph**	**Generated text**
Property understanding	(*Ciudad Ayala*, *government type*, *Council-manager government*), (*Ciudad Ayala*, *population density*, *1604.0*), (*Ciudad Ayala*, *time zone*, *Pacific Daylight Time*), (*Ciudad Ayala*, *time zone*, *Pacific Standard Time Zone*), (*Ciudad Ayala*, *leader*, *Juan Nolasco*)	Лидером Чикаго Аялах является муниципальный правительство, который управляется советом и имеет в качестве лидера Хусейна Нолоссо. Кроме того, в этом районе можно найти длину населенных пунктов 1604,0, а также в настоящее время в этом районе - Международный день-лидер и Период Пасифик. *Translation:* The leader of Chicago Ayalah is a council-managed municipal government that has Hussein Nolosso as its leader. You can also find the length of human settlements is 1604.0 in the area, and currently the International Day Leader and Pacific Period are in the area.
Topic change	(*Jung Jin-sun*, *height*, *+185 centimetre*)	Высота Юджина Ян-сана составляет - 185 метров. *Translation:* The elevation of Eugene Yang-san is -185 metres. *Explanation:* While the word высота in Russian literally means "height", it is not applied to people, only to buildings and geographical objects, like mountains.
Complex text	(*Uddheden*, *instance of*, *Urban areas in Sweden*), (*Uddheden*, *country*, *Sweden*), (*Uddheden*, *located in area*, *Sunne Municipality*)	Uddheden находится в районе Sunne Municipality в Швеции и является одним из примеров городских районов в Швеции. *Translation:* Uddheden is located in Sunne Municipality in Sweden and is one example of an urban neighbourhood in Sweden.
Garbage	(*Bootleg Series Volume 1: The Quine Tapes*, *preceded by*, *Squeeze (The Velvet Underground album)*), (*Bootleg Series Volume 1: The Quine Tapes*, *record label*, *Polydor Records*), (*Bootleg Series Volume 1: The Quine Tapes*, *record label*, *Universal Music Group*)	"Пьетро-во-

Table 10: Error annotation examples for Russian: property understanding, topic change, complex text and garbage

	#	A	O	R	ED	PU	TC	CT	garbage	good
Results for each model										
CP	30	0.13	0.2	0.3	0.07	0.07	0.0	0.6	0.0	0.2
CunI	40	0.95	0.95	0.25	0.8	0.28	0.45	0.0	0.05	0.0
T5$_{ft}$	30	0.27	0.33	0.03	0.0	0.23	0.03	0.43	0.0	0.23
Interno	40	0.57	0.6	0.32	0.72	0.22	0.1	0.15	0.2	0.0
mT0$_{ft}$	40	0.85	0.85	0.48	0.78	0.2	0.2	0.0	0.15	0.0
mT5$_{base,ft}$	70	0.51	0.51	0.34	0.44	0.23	0.2	0.01	0.47	0.0
mT5$_{base,ft}$ru	40	0.72	0.72	0.42	0.68	0.32	0.32	0.0	0.28	0.0
mT5$_{base,ft}$en	30	0.23	0.23	0.23	0.13	0.1	0.03	0.03	0.73	0.0
mT5$_{large,ft}$	70	0.5	0.49	0.3	0.31	0.13	0.21	0.01	0.49	0.0
mT5$_{large,ft}$ru	40	0.5	0.5	0.32	0.35	0.12	0.38	0.0	0.5	0.0
mT5$_{large,ft}$en	30	0.5	0.47	0.27	0.27	0.13	0.0	0.03	0.47	0.0
Results for each dataset										
KELM-E en	20	0.1	0.1	0.1	0.05	0.1	0.0	0.05	0.4	0.45
KELM-E ru	25	0.52	0.56	0.16	0.56	0.2	0.2	0.0	0.44	0.0
KELM-E+P en	100	0.32	0.35	0.23	0.13	0.14	0.02	0.32	0.28	0.04
KELM-E+P en (0.5, 0.6)	20	0.35	0.35	0.25	0.15	0.25	0.0	0.45	0.1	0.1
KELM-E+P en (0.6, 0.7)	20	0.35	0.4	0.15	0.2	0.0	0.1	0.35	0.2	0.0
KELM-E+P en (0.7, 0.8)	20	0.5	0.5	0.35	0.15	0.25	0.0	0.2	0.25	0.05
KELM-E+P en (0.8, 0.9)	20	0.25	0.35	0.3	0.1	0.15	0.0	0.2	0.4	0.05
KELM-E+P en (0.9, 1.0)	20	0.15	0.15	0.1	0.05	0.05	0.0	0.4	0.45	0.0
KELM-E+P ru	125	0.69	0.68	0.38	0.62	0.25	0.33	0.02	0.28	0.0
KELM-E+P ru (0.5, 0.6)	25	0.88	0.84	0.12	0.88	0.4	0.28	0.04	0.08	0.0
KELM-E+P ru (0.6, 0.7)	25	0.8	0.8	0.44	0.76	0.4	0.36	0.04	0.12	0.0
KELM-E+P ru (0.7, 0.8)	25	0.64	0.64	0.48	0.56	0.24	0.36	0.0	0.36	0.0
KELM-E+P ru (0.8, 0.9)	25	0.52	0.52	0.4	0.48	0.08	0.2	0.04	0.44	0.0
KELM-E+P ru (0.9, 1.0)	25	0.6	0.6	0.44	0.44	0.12	0.44	0.0	0.4	0.0
ru-WebNLG C	25	0.88	0.92	0.76	0.64	0.04	0.36	0.04	0.04	0.0
ru-WebNLG E	25	0.92	0.92	0.08	1.0	0.36	0.12	0.08	0.0	0.0
Results for each language										
English	120	0.28	0.31	0.21	0.12	0.13	0.02	0.28	0.3	0.11
Russian	200	0.72	0.72	0.36	0.66	0.23	0.29	0.03	0.24	0.0

Table 11: Full results of qualitative analysis. A: addition, O: omission, R: repetition, ED: entity distortion, PU: property understanding error, TC: topic change, CT: complex text. Ratio – yes / (# of instances in the category). The # column – number of instances to annotate in the category