The 2023 WebNLG Shared Task on Low Resource Languages Overview and Evaluation Results (WebNLG 2023)

Liam Cripwell[3,4], **Anya Belz**[1,6], **Claire Gardent**[3], **Albert Gatt**[5], **Claudia Borg**[2], **Marthese Borg**[2], **John Judge**[1], **Michela Lorandi**[1], **Anna Nikiforovskaya**[3,4], **William Soto-Martinez**[3,4], **Craig Thomson**[6]

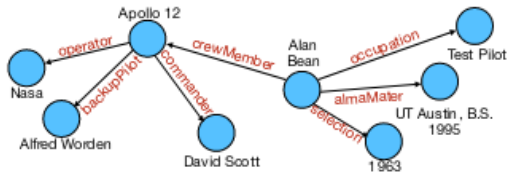
[1]ADAPT/DCU, Ireland; [2]University of Malta, Malta; [3]CNRS/LORIA, France; [4]Université de Lorraine, France; [5]Utrecht University, the Netherlands; [6]University of Aberdeen, UK

Abstract

The WebNLG task consists of mapping a knowledge graph to a text verbalising the content of that graph. The 2017 WebNLG edition required participating systems to generate English text from a set of DBpedia triples, while the 2020 WebNLG+ challenge additionally included generation into Russian and semantic parsing of English and Russian texts. In contrast, WebNLG 2023 focuses on four under-resourced languages which are severely under-represented in research on text generation, namely Breton, Irish, Maltese and Welsh. In addition, WebNLG 2023 once again includes Russian. In this paper, we present the organisation of the shared task (data, timeline, evaluation), briefly describe the participating systems and summarise results for participating systems.

1 Introduction

The WebNLG challenges seeks to add to the research on Knowledge Graph verbalisation i.e., how to convert a knowledge graph into a text verbalising its content. This is illustrated in Figure 1 where the text shown conveys the content of the input graph.



Alan Bean graduated from UT Austin in 1955 with a Bachelor of Science degree. He was hired by NASA in 1963 and served as a test pilot. Apollo 12's backup pilot was Alfred Worden and was commanded by David Scott.

Figure 1: WebNLG Input/Output Example: the generated text should convey the content of the input graph.

The first edition of the challenge, the 2017 WebNLG shared task, required participating systems to generate English text from a set of DBpedia triples (Gardent et al., 2017). The 2020 WebNLG+ challenge encompassed four tasks: RDF-to-English, RDF-to-Russian, English-to-RDF and Russian-to-RDF.

With the development of large-scale pretrained models, research in automatic text generation has acquired new impetus. Yet, the current state-of-the-art is dominated by a handful of languages, for which training data is relatively easy to acquire. At the same time, the field has recently witnessed some encouraging developments which focus on generation for under-resourced and under-represented languages (Abhishek et al., 2022; Team et al., 2022). This trend is paralleled by a growing interest in multilingual models and applications in NLP more broadly.

The WebNLG 2023 Challenge was organised in response to these trends and specifically addresses generation for four under-resourced languages: Breton, Irish, Maltese and Welsh. In addition, WebNLG 2023 once again includes Russian.

Timeline. Noisy training data, gold development data and evaluation scripts were released on February 24, 2023. The test data was made available on June 8th and the deadline for submitting system results was June 15th. Automatic evaluation results were announced on June 13th and the human evaluation results on August 18th. Results were first released anonymously so that participants had the opportunity to withdraw their systems.

In what follows, we summarise the main features of WebNLG 2023. Section 2 describes the datasets used for the challenge. Section 3 briefly presents the participating systems (more detailed descriptions are provided by the participants in separate papers). Section 4 introduces the evaluation methodology, Section 5 discusses the automatic evaluation results for participating systems and Section 6 the human evaluation results. Finally, Sec-

Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023), pages 55–66

Language	Nb. Train Items	Nb. Dev Items	Nb. Test Items
Breton (br)	13,211	1,399	1,778
Welsh (cy)	13,211	1,665	1,778
Irish (ga)	13,211	1,665	1,778
Maltese (mt)	13,211	1,665	1,778
Russian (ru)	5,573	790	1,101

Table 1: Data statistics for each of the supported languages. Training data is 'noisy', i.e. automatically translated; Dev and Test data are translations by professionals of the WebNLG 2020 dev and test English texts .

tion 7 reports correlations between automatic and human-assessed evaluation measures. Section 8 concludes with broad findings and pointers to future developments.

2 Data

To obtain the development and test data for each of the low-resource languages (Breton, Irish, Maltese, and Welsh), we had professional translators manually translate the English text from the WebNLG 2020 development and test sets, given both the English text and the input RDF graph.[1] We only consider the first reference of each test example in the original English dataset when translating, except in the case of Breton which contains 2 translated references for some of the test items. For Russian, we continue to use the WebNLG$^+$ 2020 data.

In addition to the professionally translated dev and test data, we provide optional 'noisy' training data for the low-resource target languages, which was obtained via machine translation of the texts in the English WebNLG 2020 training set. For this, we used the 24-layer Zero machine translation system (Zhang et al., 2020a).[2]

Table 1 provides a summary of the data provided for each language.

3 Participating Systems

This section provides a brief overview of the participating systems and of the runs submitted by the participants to the shared task. Further results obtained after the submission deadline and more details about each systems are available in the system descriptions provided by the participants (Hari et al., 2023; Kazakov et al., 2023; Kumar et al., 2023; Lorandi and Belz, 2023; Mille et al., 2023).

Seven teams submitted a system run and two teams withdrew after the automatic evaluation results for test set were shared with participants, yielding a final total of five participating teams (Table 2). Of the 5 runs submitted, only one (CUNI-Wue) included output for all languages. Two teams concentrated on a single language (DCU/TCD-FORGe on Irish and Interno on Russian). None of the remaining two participants submitted output for Breton, with one participant (IREL) submitting output for all languages except Breton and the other (DCU-NLG-PBN) for only three languages (Irish, Welsh, Maltese).

In terms of architecture, monolingual submissions were either rule-based (Mille et al., 2023) or a monolingual LLM fine-tuned on the WebNLG data (Hari et al., 2023). Submissions targeting multiple languages predominantly adopted an NLG+MT pipeline approach.

3.1 Monolingual Models

Interno (Kazakov et al., 2023) focuses on Russian and uses the pretrained large language model FRED-T5 (Full-scale Russian Enhanced Denoisers T5[3], 1,700 M Parameters) fine-tuned on the WebNLG 2020 training dataset to convert RDF graphs into Russian texts. Various prompts are experimented with which either consist of the input graphs or of graphs enriched with translation information for entities and relations. Results indicate that the translation data fail to significantly improve results. The system code is available at https://github.com/Ivan30003/webnlg_interno.

DCU/TCD-FORGe (Mille et al., 2023) converts RDF graphs into Irish using a rule-based pipeline consisting of four components: triple lexicalisation, generation of non-inflected Irish text, inflection generation, and post-processing. The pipeline is available at https://github.com/mille-s/DCU_TCD-FORGe_WebNLG23.

[1] Due to translation constraints, the Breton development set only contains 1,399 entries from the WebNLG 2020 English dev set. These entries were randomly sampled while making sure to cover texts from all triple set size groups.

[2] https://github.com/bzhangGo/zero/tree/master/docs/multilingual_laln_lalt

[3] https://huggingface.co/ai-forever/FRED-T5-1.7B

Team	Affiliation	Country	Breton	Welsh	Irish	Maltese	Russian
CUNI-Wue	Charles University	Czechia	✓	✓	✓	✓	✓
DCU/TCD-FORGe	ADAPT/DCU/Trinity College	Ireland	-	-	✓	-	-
Interno	Pulkovo Observatory	Russia	-	-	-	-	✓
IREL	IIT Hyderabad	India		✓	✓	✓	✓
DCU-NLG-PBN	ADAPT/DCU	Ireland	-	✓	✓	✓	-

Table 2: WebNLG 2023 Participants.

3.2 NLG+MT Models

IREL (Hari et al., 2023) adopts an NLG+MT approach. RDF graphs are first translated into English using T5-small, fine-tuned on the training split of the WebNLG 2020 dataset for English. The generated English text is then translated to Irish, Maltese, Russian and Welsh using the distilled variant of NLLB (Team et al., 2022). As NLLB does not handle Breton, the approach is not applied to the Breton data. The training code and model checkpoints are available at https://shorturl.at/hsNO4.

CUNI-Wue (Kumar et al., 2023) also uses an NLG+MT approach using improved training data, custom decoding and multi-tasking.

As for the Shared Task baseline, the input to MT is the English text generated by the best WebNLG 2020 model (**?**); the MT system is Zero MT..

Training data for machine translation from English into the Shared Task languages is created as follows. For Maltese, Irish and Welsh, the NLLB MT system is applied to the English texts of WebNLG 2020 (instead of Edinburgh's Zero MT model) while for Breton (which is not handled by NLLB), the translations produced by the Zero MT model for this same data are filtered using a length-based heuristic designed to identify incomplete translations. For Russian, the training data is the data from WebNLG 2020.

The English texts generated by the best WebNLG 2020 model are then converted into Maltese, Irish, Welsh, Russian and Breton using mT5 fine-tuned on the parallel data created from the WebNLG 2020 English texts.

Two further refinements are experimented with: custom decoding (split-and-generate) and multi-task learning. For custom decoding, the input RDF graphs are partitioned and the texts generated from each partition subset are concatenated to produce the final output. Multi-tasking includes, in addition to the RDF-to-Text generation main task, translation from English and RDF-to-English text as auxiliary tasks. The model learns to distinguish tasks by different prompts.

For Maltese, Irish and Welsh the submitted variants are multi-task learning + split-and-generate; for Breton data filtering + split-and-generate; and for Russian a multilingual setup without split-and-generate. The code and submission outputs are available at https://github.com/knalin55/CUNI_Wue-WebNLG23_Submission.

3.3 Very Large Language Models, no training

DCU-NLG-PBN (Lorandi and Belz, 2023) experimented with very large language models (GPT-3.5 and 4) without training or finetuning of any kind, testing a range of prompt types and formats on a small sample of example input/output pairs and evaluating the two most promising prompts in two scenarios: (i) direct generation into the under-resourced language, and (ii) generation into English followed by translation into the under-resourced language. The variant submitted to the WebNLG 2023 shared task is the few-shot + translation system variant. All code and results are available at https://github.com/DCU-NLG/DCU-NLG-PBN.

4 Evaluation Methodology

4.1 Automatic Metrics

The participating systems were automatically evaluated with some of the most popular text generation metrics. Specifically, we considered BLEU (Papineni et al., 2002), TER (Snover et al., 2006), chrF++ (Popović, 2017) (with word bigrams, character 6-grams and $\beta = 2$), and BERTScore (Zhang et al., 2020b)[4]. We use the SacreBLEU implementation of BLEU[5] (Post, 2018), the pyter implementation of TER[6], and the official implementations of chrF++[7] and BERTScore[8].

[4]We compute BERTScore for all languages except Maltese, as it is not supported.

[5]https://github.com/mjpost/sacrebleu

[6]https://pypi.org/project/pyter/

[7]https://github.com/m-popovic/chrF/blob/master/chrF%2B%2B.py

[8]https://github.com/google-research/bert/tree/master

All languages except Breton and Russian contain only a single reference for each test instance, and so only these two languages were evaluated in a multi-reference scenario. Each Breton hypothesis was compared with up to 2 references, and each Russian one with up to 7 references. On average, Breton data has 1.28 references per test instance, and Russian data has 2.52 references per instance. We tokenised the texts using `razdel`[9] for Russian and the NLTK framework (Bird et al., 2009) for the other languages (BERTScore uses its own tokenizer, however).

As a baseline system for the new languages, we use the highest performing RDF-to-English system from the 2020 challenge (Amazon (Guo et al., 2020)) and automatically translate its outputs into the target language using Edinburgh's Zero multilingual MT model (Zhang et al., 2020a). For Russian, we use the baseline (FORGE2020) and highest performing system (CUNI-Ufal (Kasner and Dušek, 2020)) from the 2020 challenge.

4.2 Human Evaluation

The WebNLG 2023 human evaluations assessed system outputs in terms of the quality criteria of Fluency, Absence of Omissions, Absence of Additions, and Absence of Unnecessary Repetition. These were defined and explained to evaluators as can be seen in the instructions document; instructions and other details and resources of the evaluations were published as a preregistration bundle on 25 July 2023.[10]

The same evaluations were carried out for each language except Breton for which only one system was submitted. Once the design was complete, we conducted a pilot evaluation on a set of 10 English WebNLG input/output pairs with six of the authors who were not directly involved in developing the evaluation design. Fleiss's kappa values in the pilot were as follows:

Fleiss's κ for human evaluation methods			
Fluency	−Omissions	−Additions	−Repetitions
0.216	0.908	0.81	0.811

Kappa for Absence of Omissions, Absence of Additions, and Absence of Unnecessary Repetition is high. The lower Fluency kappa is not entirely surprising, as evaluators often disagree on what makes a text (not) fluent. However, as we will see below, kappa was not higher among the authors than among the evaluators which *is* surprising.

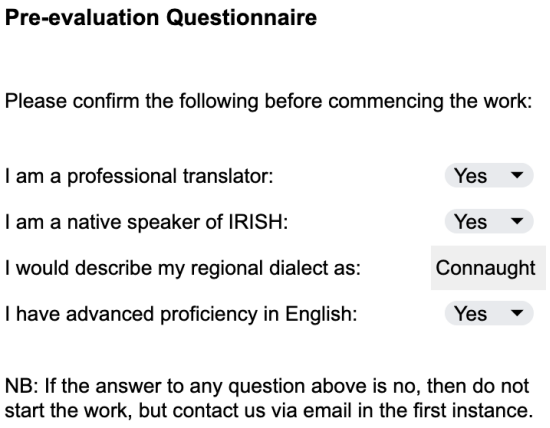


Figure 2: Exclusion questions and dialect check asked of evaluators (here completed by an Irish evaluator).

The main steps in conducting the evaluation for each language were the following:

1. Recruitment of professional translators via translation agencies.
2. Online training session for evaluators using a Google spreadsheet (Figure 3) and documents and the same 10 English outputs, followed by feedback if the overall kappa for any of the criteria dropped to below 0.7 after the addition of an evaluator's scores to the pilot results.
3. Full evaluation of 100 outputs in the human evaluation test set (see below).
4. Aggregation and analysis of results as described below.

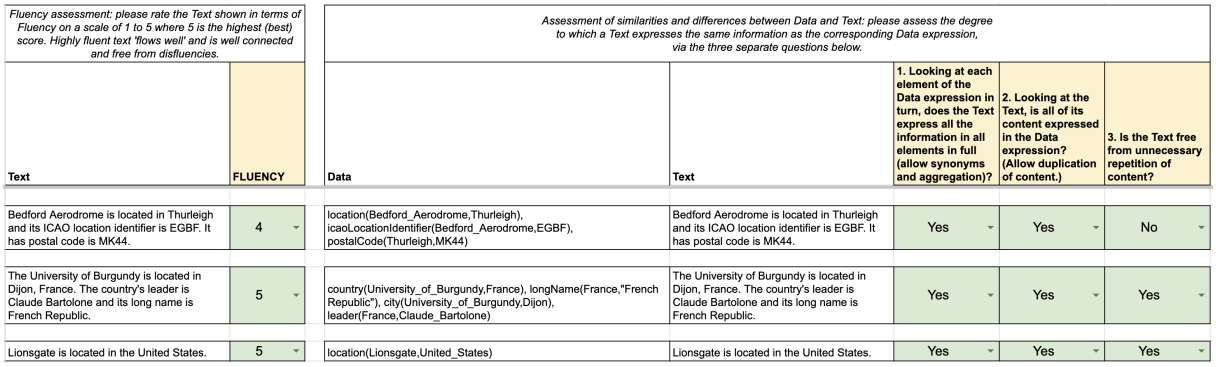
Before commencing the evaluation, translators were asked the questions in Figure 2.

Human evaluation test data: We randomly selected 100 inputs and corresponding system outputs plus human reference text (from the test data itself). Selection was performed with stratification for WebNLG category and number of triples in the input, with the same inputs being used for all low-resource languages (Russian used a different sample of items because of its differing test set).

Allocation of items to evaluators: We used a Repeated Latin Squares design which ensures that each evaluator sees the same number of outputs from each system and for each test set item. For Irish, where we had 5 systems, there were twenty

[9] https://github.com/natasha/razdel

[10] Summary on AsPredicted #139263: https://aspredicted.org/~Mgcdw2J2h6; full details and resources: https://github.com/nlgcat/webnlg2023_human_eval_preregistration



Fluency assessment: please rate the Text shown in terms of Fluency on a scale of 1 to 5 where 5 is the highest (best) score. Highly fluent text 'flows well' and is well connected and free from disfluencies.		*Assessment of similarities and differences between Data and Text: please assess the degree to which a Text expresses the same information as the corresponding Data expression, via the three separate questions below.*				
Text	**FLUENCY**	**Data**	**Text**	**1. Looking at each element of the Data expression in turn, does the Text express all the information in all elements in full (allow synonyms and aggregation)?**	**2. Looking at the Text, is all of its content expressed in the Data expression? (Allow duplication of content.)**	**3. Is the Text free from unnecessary repetition of content?**
Bedford Aerodrome is located in Thurleigh and its ICAO location identifier is EGBF. It has postal code is MK44.	4	location(Bedford_Aerodrome,Thurleigh), icaoLocationIdentifier(Bedford_Aerodrome,EGBF), postalCode(Thurleigh,MK44)	Bedford Aerodrome is located in Thurleigh and its ICAO location identifier is EGBF. It has postal code is MK44.	Yes	Yes	No
The University of Burgundy is located in Dijon, France. The country's leader is Claude Bartolone and its long name is French Republic.	5	country(University_of_Burgundy,France), longName(France,"French Republic"), city(University_of_Burgundy,Dijon), leader(France,Claude_Bartolone)	The University of Burgundy is located in Dijon, France. The country's leader is Claude Bartolone and its long name is French Republic.	Yes	Yes	Yes
Lionsgate is located in the United States.	5	location(Lionsgate,United_States)	Lionsgate is located in the United States.	Yes	Yes	Yes

Figure 3: Human evaluation: Screen shot with items from pilot evaluation with English outputs.

5 × 5 squares, and 2,000 individual judgements (5 evaluators × 4 criteria × 100 outputs), and 400 per system. For the other languages (4 systems) there were twenty-five 4 × 4 squares, and 1,600 individual judgements (4 evaluators × 4 criteria × 100 outputs), again 400 per system.

Aggregation: We used the Fluency assessments unchanged, computing the mean over scores. For the other three criteria, we computed the proportions of Yes scores (equal to the mean of scores mapped Yes=1, No=0).

Analysis: For each language, we carried out four univariate ANOVAs with System as the fixed factor, and Fluency (1–5 ratings), Omission (Yes/No assessments), Addition (Yes/No assessments), and Repetition (Yes/No assessments) as the dependent variables each in one of the ANOVAs. In the results section below, we report F-ratios and their statistical significance, and the homogeneous subsets of systems as determined by a post-hoc Tukey HSD analysis, reflecting significant pairwise differences between systems. The results from the latter are shown, alongside the mean assessment values, in tables where systems whose scores are not significantly different (at the .05 level) share a letter.

We also report Pearson and Spearman correlations with probabilities of statistical significance, between individual Fluency, Omissions, Additions and Repetitions assessments.

5 Results of Automatic Evaluation

In this section, we present results of the evaluation using automatic metrics. These are summarised in Table 3 for each of the four languages under consideration, ordered by chrF++ score. All systems outperform the baselines (or one of the baselines, in the case of Russian), on at least some of the metrics.

As regards the monolingual systems, Interno is ranked (joint) first for Russian on all metric, while DCU/TCD-FORGe is ranked second for BLEU and chrF++, fourth for TER, and third or fourth for BERT P/R/F1. While this suggests that monolingual approaches (whether LLM-based or rule-based) can provide competitive solutions for specific languages, the broader pattern across the four languages shows that combining LLMs with machine translation is more effective in the absence of large amounts of training data.

DCU-NLG-PBN which combines GPT-3.5 with Google Translate and is not trained/finetuned on any WebNLG data, outperforms all systems on all metrics in the three languages it was tested on. It is also instructive to consider the relative performance of IREL and CUNI-Wue, the latter leveraging MT for improved training data, while the former uses it to translate outputs. IREL narrowly outperforms CUNI-Wue on Maltese, Welsh and Irish on all metrics except BLEU on Irish. In summary, these results show a continued role for MT to handle under-resourced languages, even with very large models, as in the case of DCU-NLG-PBN, who report worse performance with direct generation (Lorandi and Belz, 2023).

The inclusion of Russian in WebNLG 2023 provides a point of comparison with the results for the same language in the 2020 edition (Castro Ferreira et al., 2020). In 2020, the CUNI-Ufal system Kasner and Dušek (2020) ranked second on Russian. This year, it is narrowly outperformed by Interno and CUNI-Wue, at least on BLEU and chrF++. On BERTScore, the systems are very close. Indeed, results on Russian for the present edition provide very small improvements over the best results for 2020.

	BLEU	chrF++	TER ↓
DCU-NLG-PBN	21.27	0.52	0.65
IREL	16.49	0.47	0.7
CUNI-Wue	14.02	0.45	0.78
Amazon+Zero	15.60	0.42	0.67

(a) Maltese

	BLEU	chrF++	TER ↓	BERT_P	BERT_R	BERT_F1
Amazon+Zero	9.92	0.33	0.76	0.77	0.73	0.75
CUNI-Wue	10.09	0.33	0.80	0.76	0.73	0.74

(b) Breton

	BLEU	chrF++	TER ↓	BERT_P	BERT_R	BERT_F1
DCU-NLG-PBN	25.11	0.55	0.64	0.83	0.83	0.83
IREL	20.97	0.49	0.67	0.82	0.8	0.81
CUNI-Wue	17.00	0.45	0.79	0.79	0.78	0.79
Amazon+Zero	10.70	0.36	0.77	0.78	0.75	0.76

(c) Welsh

	BLEU	chrF++	TER ↓	BERT_P	BERT_R	BERT_F1
DCU-NLG-PBN	20.40	0.51	0.69	0.81	0.8	0.81
DCU/TCD-FORGe	16.66	0.44	0.75	0.79	0.76	0.77
IREL	15.66	0.44	0.73	0.8	0.77	0.78
CUNI-Wue	15.87	0.43	0.78	0.78	0.77	0.77
Amazon+Zero	11.63	0.36	0.74	0.78	0.74	0.76

(d) Irish

	BLEU	chrF++	TER ↓	BERT_P	BERT_R	BERT_F1
Interno	54.68	0.69	0.37	0.92	0.91	0.92
CUNI-Wue	54.52	0.69	0.38	0.92	0.91	0.91
CUNI-Ufal (2020)	52.9	0.68	0.40	0.91	0.91	0.91
IREL	36.01	0.57	0.53	0.88	0.87	0.87
FORGE (2020)	25.5	0.51	0.67	0.84	0.84	0.84

(e) Russian

Table 3: Results of the automatic evaluation, per language. For each language, results are in descending order of chrF++ score. Baseline results are shaded in light grey: for Russian, we include results from the 2020 version of FORGe (Mille et al., 2019) and for CUNI-Ufal (Kasner and Dušek, 2020), both reported in WebNLG 2020 (Castro Ferreira et al., 2020). Note that BERTScore is not available for Maltese.

6 Results of Human Evaluation

In this section, we present and discuss results from the human evaluations for Irish, Maltese and Welsh. For Russian, we were unable to recruit and train translators able to perform the task reliably, as discussed in the next section below, and in Appendix A.

6.1 Quality assurance checks

For Irish, Maltese and Breton, most evaluators passed the test in the training session easily. One evaluator for Maltese and one for Welsh did not initially pass the threshold. The Welsh evaluator passed the threshold after additional explanation. The Maltese evaluator failed the quality threshold on two of the criteria, and despite prolonged exchange did not manage to pass it, and had to be replaced. The replacement also required repeated additional explanation. All but one evaluator in these languages increased Fleiss's kappa for Fluency when added to the pilot pool of evaluators. On average, compared to the table of pilot kappas (Section 4.2), evaluators increased Fluency to 0.2525, decreased Omissions to 0.855, and slightly decreased Additions and Repetitions to 0.7975 and 0.79, respectively.

The picture was different for Russian, where

three evaluators dropped out after completing the training session, and 3/4 of evaluators who didn't drop out, then failed the quality threshold, two on two criteria, and one on three. Feedback did not succeed in improving kappa; given our recruitment difficulties, we did not have time to replace as many as three evaluators. Concerns from the training session results were corroborated by three further sets of findings for Russian: (a) the effect sizes as per the ANOVAs reported in the appendix for System on scores are very small, and none are significant; (b) there was no correlation between the automatic metrics and the human evaluation measures at all for any measures (Table 9); and (c) there was just a single pairwise significant difference between systems for Russian (Fluency for Interno was better than for IREL), and the reference texts were not significantly better than any system on any scores, and were ranked top only for Absence of Repetitions. The human evaluation results for Russian were therefore not deemed sufficiently reliable, and are presented, for reference, in Appendix A.

6.2 Other checks

Given the overlap between submitting teams and organisers, we took several measures to address conflict of interest, and ensure independence of evaluation. One, we are making all system outputs, original human evaluation scores and scripts, and automatic evaluation scripts available, so that all results can be verified easily. Two, systems and teams were completely obscured in the evaluator spreadsheets. Three, an independent organiser double-checked that the system outputs in the spreadsheets were unchanged compared to original team submission. And four, the same independent organiser also double-checked the edit histories of the evaluator spreadsheets to ensure that any edits were either (a) done outside of the evaluation cells for the purpose of formatting and getting spreadsheets ready, or (b) done inside the evaluation cells by the evaluators only.

6.3 ANOVA and Tukey HSD results

The results of the full evaluations for Irish, Maltese and Welsh can be seen in Table 4. For each language, systems are shown in order of Fluency means, where the human-written reference texts are always ranked top. The means for the other output quality criteria mostly follow the same order, except that DCU-NLG-PBN takes top spot for Additions, Omissions and Repetitions for Maltese and Irish, and Repetitions ranks vary a lot.

The columns containing single capital letters in Table 4 show the homogeneous subsets of systems as determined by a post-hoc Tukey HSD analysis. Systems whose scores are not significantly different (at the .05 level) share a letter.

For **Fluency**, the DCU-NLG-PBN system is in second place after the reference, followed by IREL and then CUNI-Wue, for all three languages. For Irish we also have DCU/TCD-FORGe which comes in at rank 3. For **Absence of Additions**, the references and DCU-NLG-PBN are significantly better than the other two systems for Welsh and Maltese, for the latter IREL is also significantly better than CUNI-Wue; for Irish, the references, DCU-NLG-PBN and DCU/TCD-FORGe are all significantly better than CUNI-Wue; interestingly the references are significantly better only than the lowest-ranked system CUNI-Wue.

For **Absence of Omissions**, the references and DCU-NLG-PBN are significantly better than the other two systems for Welsh and Maltese. The references, DCU-NLG-PBN and DCU/TCD-FORGe are all significantly better than the other two systems for Irish.

For **Absence of Repetitions**, no significant differences were found, and results are generally very close together, for Welsh and Irish. For Maltese, the references and DCU-NLG-PBN were significantly better than CUNI-Wue. From the individual scores it looks as if repetition was simply too sporadic to yield meaningful results.

Table 5 shows the effect sizes and associated p-values for the three languages. The general picture is that for Irish, Maltese and Welsh, there is a strong and highly significant effect of *System* on output quality for all measures, except Repetition, presumably for the reasons discussed above. There are negligible and mostly non-significant effects of *Category* and *Triples* on performance. Finally, there is in all cases except Fluency a small to medium, and mostly significant, effect of *Evaluator* on performance. For Fluency, the effect of Evaluator is very substantial, for the reasons discussed previously.

The correlations between the four human-assessed quality criteria, and between these and the automatic metrics are shown in Tables 6 to 8, and are discussed in a separate section below.

Language	System	Fluency				Addition				Omission				Repetition			
Welsh	Human reference	**3.28**	A			**0.9**	A			**0.84**	A			**0.95**	A		
	DCU-NLG-PBN	3.25	A			0.86	A			0.77	A			0.94	A		
	IREL	2.67		B		0.6		B		0.47		B		0.94	A		
	CUNI-Wue	2.35		B		0.45		B		0.33		B		0.88	A		
Maltese	Human reference	**4.27**	A			0.89	A			0.85	A			0.91	A		
	DCU-NLG-PBN	4.06	A	B		**0.91**	A			**0.86**	A			**0.94**	A		
	IREL	3.74		B		0.69		B		0.56		B		0.87	A	B	
	CUNI-Wue	3.34			C	0.52			C	0.46		B		0.76		B	
Irish	Human reference	**4.07**	A			0.81	A			0.82	A			0.96	A		
	DCU-NLG-PBN	3.83	A	B		0.83	A			**0.85**	A			**0.97**	A		
	IREL	3.39		B	C	0.65	A	B		0.58		B		0.94	A		
	DCU/TCD-FORGe	3.35			C	**0.84**	A			0.81	A			0.89	A		
	CUNI-Wue	2.98			C	0.55		B		0.51		B		0.92	A		

Table 4: Post-hoc Tukey HSD (alpha = .05) analysis results for Welsh, Maltese and Irish. The columns containing single capital letters show the homogeneous subsets of systems. Human reference results are shaded in light grey.

7 Correlation between Automatic and Human Evaluation Metrics

Tables 6 to 8 are Pearson correlation matrices for all automatic and human-assessed output quality measures. The colour coding goes from vivid green for strong positive correlations to vivid red for strong negative correlations, with both colours growing paler toward 0 (no correlation).

As can be seen very clearly, there are strong correlations between all automatic metrics (note that TER is the only metric where a lower score is better). Otherwise, the strongest correlations are between Omissions and Additions for all languages. In other words, a system that omits information that is in the data from the output is also likely to add information to the output that is not in the data.

In terms of how automatic metrics on the one hand correlate with human-assessed measures on the other, the overall picture is that there are weak to medium correlations between all automatic metrics and Fluency, Additions and Omissions, but not Repetition.

The weakness of the correlations between human and automatic metrics points to an interesting side-effect of building systems for under-resourced languages, in particular when using LLMs. In the past, automatic metrics which tend to assess quality as similarity to a set of test set reference outputs were good predictors of human assessments of outputs of systems which tended to be trained on another part of the same data set. For under-resourced languages, training on reference outputs takes a back seat, hence evaluation on the basis of similarity to reference outputs is a far less effective predictor of human-assessed quality. This effect is exacerbated when using an LLM out-of-the-box as in the case of the DCU-NLG-PBN system.

8 Conclusion

Reflecting on the WebNLG 2023 results overall, it seems remarkable, especially when compared to the two previous WebNLG iterations, that DCU-NLG-PBN, a system that uses GPT-3.5 plus Google Translate in zero-shot mode, with no task-specific adaptation at all, should emerge as the overall winner: its performance is not significantly different from the human-authored reference texts on any of the human-assessed performance measures, and the (automatic metric) margins by which it outperforms systems that have been trained specifically for this task are in many cases substantial.

What is even more surprising is that in the five language/measure combinations where DCU-NLG-PBN is actually ranked higher than the human-written reference texts, the measure in question is never Fluency (which LLMs are supposed to be particularly good at), but Additions, Omissions and Repetitions (which LLMs are supposed to be particularly bad at).

Overall the shared task illustrates the limitations of current neural models when dealing with low resource languages. While fine-tuning existing encoder-decoders or decoders under-perform the NLG+MT pipeline, for languages such as Breton, where machine translation is not available or low quality, the NLG+MT approach similarly shows poor results. In both cases, the scarcity of training data restricts the quality of the generated texts. Approaches which have not been adopted by any of

Language	Fluency				Addition			
	System	Category	# Triples	Evaluator	System	Category	Triples	Evaluator
Welsh	13.64 (p<.001)	1.96 (p=.013)	2.69 (p=.014)	38.48 (p<.001)	26.12 (p<.001)	1.43 (p=.117)	1.37 (p=.226)	7.35 (p<.001)
Maltese	13.87 (p<.001)	2.64 (p<.001)	3.57 (p=.002)	21.72 (p<.001)	20.86 (p<.001)	1.69 (p=.042)	0.93 (p=.476)	2.84 (p=.038)
Irish	12.39 (p<.001)	1.23 (p=.238)	3.14 (p=.005)	50.68 (p<.001)	9.18 (p<.001)	1.1 (p=.351)	1.29 (p=.259)	9.98 (p<.001)

(a) Effect sizes for **Fluency** and **Absence of Additions**.

Language	Omission				Repetition			
	System	Category	# Triples	Evaluator	System	Category	Triples	Evaluator
Welsh	29.77 (p<.001)	1.53 (p=.081)	1.69 (p=.123)	6.0 (p<.001)	1.53 (p=.207)	2.45 (p=.001)	6.87 (p<.001)	5.88 (p<.001)
Maltese	22.05 (p<.001)	2.17 (p=.005)	2.61 (p=.017)	1.71 (p=.165)	5.66 (p<.001)	2.18 (p=.005)	2.11 (p=.052)	3.11 (p=.026)
Irish	13.22 (p<.001)	1.49 (p=.092)	1.95 (p=.071)	11.17 (p<.001)	1.73 (p=.143)	1.13 (p=.324)	2.45 (p=.024)	5.58 (p<.001)

(b) Effect sizes for **Absence of Omissions** and **Absence of Repetitions**.

Table 5: Effect sizes and their statistical significance measures ($\alpha = 0.05$) from one-way ANOVAs for Welsh, Maltese and Irish, with System, Category, Number of triples and Evaluator as factors.

Metric	BLEU	chrF++	TER ↓	Addition	Fluency	Omission	Repetition
BLEU	1.0	0.96	−0.93	0.31	0.23	0.33	0.09
chrF++	0.96	1.0	−0.89	0.36	0.27	0.37	0.09
ter ↓	−0.93	−0.89	1.0	−0.36	−0.29	−0.34	−0.15
Addition	0.31	0.36	−0.36	1.0	0.39	0.61	0.18
Fluency	0.23	0.27	−0.29	0.39	1.0	0.46	0.25
Omission	0.33	0.37	−0.34	0.61	0.46	1.0	0.13
Repetition	0.09	0.09	−0.15	0.18	0.25	0.13	1.0

Table 6: Pearson correlation matrix for all automatic and human-assessed output quality measures for Welsh.

Metric	BLEU	chrF++	TER ↓	Addition	Fluency	Omission	Repetition
BLEU	1.0	0.94	−0.93	0.14	0.25	0.17	0.06
chrF++	0.94	1.0	−0.9	0.17	0.3	0.21	0.09
TER ↓	−0.93	−0.9	1.0	−0.17	−0.28	−0.19	−0.1
Addition	0.14	0.17	−0.17	1.0	0.25	0.74	0.12
Fluency	0.25	0.3	−0.28	0.25	1.0	0.26	0.13
Omission	0.17	0.21	−0.19	0.74	0.26	1.0	0.03
Repetition	0.06	0.09	−0.1	0.12	0.13	0.03	1.0

Table 7: Pearson correlation matrix for all automatic and human-assessed output quality measures for Irish.

Metric	BLEU	chrF++	TER ↓	Addition	Fluency	Omission	Repetition
BLEU	1.0	0.95	−0.78	0.17	0.24	0.23	0.09
chrF++	0.95	1.0	−0.81	0.21	0.29	0.27	0.12
TER ↓	−0.78	−0.81	1.0	−0.25	−0.32	−0.28	−0.24
Addition	0.17	0.21	−0.25	1.0	0.41	0.62	0.29
Fluency	0.24	0.29	−0.32	0.41	1.0	0.43	0.42
Omission	0.23	0.27	−0.28	0.62	0.43	1.0	0.23
Repetition	0.09	0.12	−0.24	0.29	0.42	0.23	1.0

Table 8: Pearson correlation matrix for all automatic and human-assessed output quality measures for Maltese.

the participants but which would be interesting to explore include data augmentation, parameter efficient fine tuning and the development of languages models that are better attuned to specific language families.

The shared task also highlights the practical difficulties involved in training and testing models for low resource languages. As our human evaluation illustrates, recruiting reliable experts to evaluate system outputs is challenging. Similarly, creating development and test data is both costly and complex. We hope that the WebNLG 2023 data will encourage and foster further research on generation into low resource languages.

Acknowledgments

WebNLG 2023 is being organised under the auspices of LT-Bridge, supported by the Horizon 2020 Work Programme Spreading Excellence and Widening Participation (WIDESPREAD) 2018-2020 and the ANR funded xNLG Chair on multilingual, multi-source NLG (Gardent; award ANR-20-CHIA-0003, XNLG "Multi-lingual, Multi-Source Text Generation").

References

Tushar Abhishek, Shivprasad Sagare, Bhavyajeet Singh, Anubhav Sharma, Manish Gupta, and Vasudeva Varma. 2022. Xalign: Cross-lingual fact-to-text alignment and generation for low-resource languages. *CoRR*, abs/2202.00291.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Thiago Castro Ferreira, Claire Gardent, Nikolai Ilinykh, Chris van der Lee, Simon Mille, Diego Moussallem, and Anastasia Shimorina. 2020. The 2020 bilingual, bi-directional WebNLG+ shared task: Overview and evaluation results (WebNLG+ 2020). In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 55–76, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The WebNLG challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, Santiago de Compostela, Spain. Association for Computational Linguistics.

Qipeng Guo, Zhijing Jin, Xipeng Qiu, Weinan Zhang, David Wipf, and Zheng Zhang. 2020. CycleGT: Unsupervised graph-to-text and text-to-graph generation via cycle training. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 77–88, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Kancharla Aditya Hari, Bhavyajeet Singh, Anubhav Sharma, and Vasudeva Varma. 2023. Webnlg challenge 2023: Domain adaptive machine translation for low-resource multilingual rdf-to-text generation. Technical report, IIT Hyberhadad. WebNLG 2023 System Description.

Zdeněk Kasner and Ondřej Dušek. 2020. Train hard, finetune easy: Multilingual denoising for RDF-to-text generation. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 171–176, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Maxim Kazakov, Julia Preobrazhenskaya, Ivan Bulychev, and Aleksandr Shain. 2023. Webnlg-interno: Utilizing fred-t5 to address the rdf-to-text problem. Technical report, Pulkovo Observatory. WebNLG 2023 System Description.

Nalin Kumar, Saad Obaid ul Islam, and Ondrej Dusek. 2023. Better translation + split and generate for multilingual rdf-to-text. Technical report, Charles University. WebNLG 2023 System Description.

Michela Lorandi and Anya Belz. 2023. Data-to-text generation for severlay under-resourced languages with gpt-3.5: A bit of help needed from google translate. Technical report, ADAPT, Dublin City University. WebNLG 2023 System Description.

Simon Mille, Stamatia Dasiopoulou, Beatriz Fisas, and Leo Wanner. 2019. Teaching FORGe to verbalize DBpedia properties in Spanish. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 473–483, Tokyo, Japan. Association for Computational Linguistics.

Simon Mille, Elaine Ui Dhonnchadha, Stamatis Dasiopoulou, Lauren Cassidy, Brian Davis, and Anya Belz. 2023. Dcu/tcd-forge at webnlg'23: Irish rules! Technical report, ADAPT, Dublin City University. WebNLG 2023 System Description.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020a. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Russian Human Evaluations

As explained in the main body of the paper, we are presenting the results from the Russian human evaluation separately in this section, because of concerns about their reliability.

Table 9 is the Pearson correlation matrix for all automatic and human-assessed output quality measures for Russian. As discussed in the paper, the human and automatic measures do not correlate at all.

The results from the one-way ANOVAs for Russian can be seen in Table 10. Systems are shown in order of Fluency means. Unlike the other languages, there are three different system rankings among the four output quality measures.

The columns containing single capital letters in Table 10 show the homogeneous subsets of systems as determined by a post-hoc Tukey HSD analysis. Systems whose scores are not significantly different (at the .05 level) share a letter. Unlike the other languages, there is just one significant difference (Interno's Fluency is significantly better than IREL's.

Table 11 shows the effect sizes and associated p-values for Russian. What we should be seeing is a strong and significant effect of System on each output quality measure, but we're only seeing a slight effect on Fluency, for the other measures, there is no effect, i.e. the scores have a lot of randomness in them.

Metric	BLEU	chrF++	TER ↓	Addition	Fluency	Omission	Repetition
BLEU	1.0	0.96	−0.92	0.06	0.1	0.05	0.07
chrF++	0.96	1.0	−0.92	0.07	0.14	0.08	0.07
TER ↓	−0.92	−0.92	1.0	−0.08	−0.13	−0.06	−0.11
Addition	0.06	0.07	−0.08	1.0	0.35	0.47	0.25
Fluency	0.1	0.14	−0.13	0.35	1.0	0.11	0.37
Omission	0.05	0.08	−0.06	0.47	0.11	1.0	0.16
Repetition	0.07	0.07	−0.11	0.25	0.37	0.16	1.0

Table 9: Pearson correlation matrix for all automatic and human-assessed output quality measures for Russian.

Language	**System**	**Fluency**			**Addition**		**Omission**		**Repetition**	
Russian	Interno	4.13	A		0.83	A	0.91	A	0.9	A
	Human reference	4.02	A	B	0.82	A	0.87	A	0.92	A
	CUNI-Wue	3.99	A	B	0.85	A	0.84	A	0.89	A
	IREL	3.65		B	0.79	A	0.84	A	0.88	A

Table 10: Post-hoc Tukey HSD (alpha = .05) analysis results for Russian. The columns containing single capital letters show the homogeneous subsets of systems. Human reference results are shaded in light grey.

Language	**Fluency**				**Addition**			
	System	Category	# Triples	Evaluator	System	Category	Triples	Evaluator
Russian	3.51 (p=.015)	1.53 (p=.145)	4.39 (p<.001)	77.42 (p<.001)	0.43 (p=.735)	1.81 (p=.075)	1.41 (p=.208)	14.86 (p<.001)

(a) ANOVA results for Fluency and Addition.

Language	**Omission**				**Repetition**			
	System	Category	# Triples	Evaluator	System	Category	Triples	Evaluator
Russian	0.94 (p=.422)	3.16 (p=.002)	2.33 (p=.032)	6.74 (p<.001)	0.31 (p=.815)	1.84 (p=.069)	4.2 (p<.001)	29.32 (p<.001)

(b) ANOVA results for Omission and Repetition .

Table 11: ANOVA results for Russian based on System, Category, Number of triples and Evaluator.