Semantic Evaluation of Multilingual Knowledge Graph-to-Text Generation via NLI Fine-Tuning: Precision, Recall and F1 scores

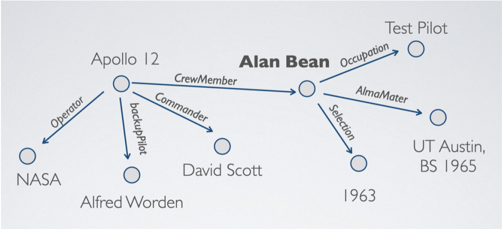
William Soto Martinez
Université de Lorraine / LORIA
william-eduardo.soto-martinez@loria.fr

Yannick Parmentier
Université de Lorraine / LORIA
yannick.parmentier@loria.fr

Claire Gardent
CNRS/LORIA and Université de Lorraine
claire.gardent@loria.fr

Abstract

Performance in the Knowledge Graph-to-Text generation has improved over time, particularly in English. However, models are still prone to mistakes like Additions and Omissions. Furthermore, few languages are taken into account since both train and test data are not readily available. In this paper, we hope to facilitate the development and improvement of multilingual KG-to-Text models by providing a multilingual evaluation framework that is referenceless and permits estimating how much a KG-to-Text Model under- (omission) or over- (addition) generates. We focus on two high (English, Russian) and five low (Breton, Irish, Maltese, Welsh, Xhosa) resource languages and show that our metric has fair to moderate correlation with reference-based metrics, positioning it as a consistent alternative when no references are available. We also show that our metric outperforms prior reference-less metrics in correlation with existing human judgments. Additional human evaluation shows moderate to strong correlation with human annotators in assessing precision and recall at a higher granularity level than shown in previous studies. Since our metric provides scores for precision and recall, it helps better assess the level of over- or under-generation of multilingual KG-to-Text models. We make our data, code and models available[1].



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: Example of a Knowledge Graph from the DB-pedia Knowledge Base together with a possible correct verbalization.

1 Introduction

Figure 1 shows an example of a Knowledge Graph and its verbalization. In a Knowledge Graph (KG), each edge represents a fact as a (subject, predicate, object) triple. To make these graphs more accessible, KG-to-Text generation models have been proposed whose function is to convert KGs into natural language. A key constraint on this task is that generation should be *semantically faithful*, meaning that the generated text should express all and only the content represented by the input KG.

While KG-to-Text generation models have steadily improved over the years both in terms of performance and of range of target languages they can handle (Gardent et al., 2017; Castro Ferreira et al., 2020; Cripwell et al., 2023), recent results indicate that semantic faithfulness is still an issue since the generated texts can either contain information not present in the input (Additions) or, conversely, fail to express all the information present in the input (Omissions). These issues are particularly prevalent when generating into under-resourced languages (Cripwell et al., 2023) or out-of-domain topics (Nikiforovskaya and Gardent, 2024).

In this paper, we provide a novel framework for the evaluation of KG-to-Text Models which, we hope, will help support the development of multilingual, semantically faithful KG-to-Text models. We make the following contributions:

1) A new reference-less multilingual metric that quantifies how much a model under- (omissions) or over- (additions) generates. This metric provides three scores: precision, recall, and F1. Intuitively, the graph acts as a reference. Hence, precision is the ratio between correct information in the text

[1] https://gitlab.inria.fr/wsotomar/semantic-evaluation-of-multilingual-data-to-text-generation-via-nli-fine-tuning

Findings of the Association for Computational Linguistics: ACL 2025, pages 10407–10427
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

<table>
<tr><th colspan="4">Graph</th></tr>
<tr><td colspan="4">Alan Bean | birthDate | 1932-03-15
Alan Bean | almaMater | UT Austin, B.S. 1955
Alan Bean | birthPlace | Wheeler, Texas</td></tr>
<tr><th>Texts</th><th>Precision</th><th>Recall</th><th>Errors</th></tr>
<tr><td>Alan Bean was born on March 15, 1932.</td><td>1/1</td><td>1/3</td><td>2O</td></tr>
<tr><td>Alan Bean was born in Wheeler, Texas and was in the Apollo 12 mission.</td><td>1/2</td><td>1/3</td><td>1A, 2O</td></tr>
<tr><td>Alan Bean was born on March 15, 1932 in Wheeler, Texas.
He received a Bachelor of Science degree at the
University of Texas at Austin in 1955.</td><td>3/3</td><td>3/3</td><td>None</td></tr>
</table>

Table 1: Example KG graph and some possible lexicalizations of this graph. The lexicalizations have different precision and recall scores as well as an explanation of the errors that cause the scores (O:Omission, A:Addition, A triple that is not fully expressed by the text counts as omitted and vice versa for additions).

and total information in the text (how much of the generated text is correct?) while recall is the ratio between correct information in the text and information in the input graph (how much of the input graph does the text convey?). F1 is their harmonic mean (See Table 1 for some illustrating examples).

2) A methodology for creating the training data necessary to train our metric.

3) Testing on both high (English, Russian) and low (Breton, Irish, Maltese, Welsh, Xhosa) resource languages, we compute correlation with both existing reference-based metrics and human judgments. We show that correlation with reference-based metrics is fair to moderate, which indicates that our metric, although reference-less, can be used to a certain extent in place of reference-based metrics; particularly when references are not available. When comparing with human judgments, we show that correlation with our metric outperforms the correlation obtained on the same data by other existing reference-less metrics developed for English KG-to-Text like Data-QuestEval (Rebuffel et al., 2021) and FactSpotter (Zhang et al., 2023).

2 Related Work

Zero-shot NLI Classification for Semantic Accuracy. Dušek and Kasner (2020) proposed to evaluate the semantic accuracy of English KG-to-Text generation by leveraging the zero-shot abilities of the English-based RoBERTa-Large-NLI model (Liu et al., 2019)[2]. In their work, they try two approaches: one to search for Omissions and one to search for Additions. In the first one, they use the entire generated text as a premise and iterate over every individual fact from the input graph as a hypothesis, marking facts not classified as entailments as Omissions. In the second approach, they use the entire input graph as a premise and the generated text as a hypothesis, marking the presence of Additions when the text is not classified as entailment. This approach is tested exclusively in English and, while it provides high granularity when measuring Omissions, it is less specific when measuring Additions.

Fine-tuning on Synthetic Data for Factual Faithfulness. Zhang et al. (2023) went a step further and fine-tuned an English model, first an Electra-Base-Discriminator (Clark et al., 2020)[3] then a DeBERTa-V3-Base (He et al., 2021)[4], on synthetic data to detect whether a given fact is present in a generated text (akin to Dušek and Kasner's Omission check). Their training data consist of real positive (Text, Fact) pairs and synthetic negative (Text, Fact) pairs made of 90% type I errors (where they perturbed the fact by changing its subject, predicate, and/or object) and 10% type II errors (where they perturbed the text by removing one or both entities from the fact and/or the n-grams most similar to the predicate). Like FactSpotter, this approach was tested exclusively in English and, while it has a high granularity when measuring Omissions, it does not directly address Additions.

Beyond KG-to-Text: Bidirectional entailment has been tried as a way of evaluating summarization. Kane et al. (2020) and Zhang and Perez-Beltrachini (2024) proposed reference-based approaches, while Chen and Eger (2023) proposed a reference-less method.

[2] https://huggingface.co/FacebookAI/roberta-large-mnli

[3] https://huggingface.co/google/electra-base-discriminator

[4] https://huggingface.co/microsoft/deberta-v3-base

3 Method

To learn our metric, we fine-tune an existing multilingual Natural Language Inference (NLI) model by adjusting its classification head to perform regression instead. We then train on data created to capture different combinations of precision and recall using Binary Cross Entropy (BCE) loss.

Given a premise and a hypothesis, NLI models predict if the hypothesis is entailed, neutral, or contradicted by the premise. For precision, we check if the text is entailed by the graph (how much of the text can be inferred from the graph). For recall we check if the graph is entailed by the text (how much of the graph content can be inferred from the text).

We are not interested in the 3 classes from the NLI head, only in the strength of the entailment between premise and hypothesis. We fine-tune the NLI classifier as a regression model by focusing only on the entailment weights from the classification head instead of the three existing output classes. We train simultaneously for precision and recall by swapping the graph and text order and targeting the respective score.

The F1 score is computed as usual (Equation 1) by taking the harmonic mean of Precision (p) and Recall (r). This score functions as a high-level proxy for semantic faithfulness: the higher the F1 score, the higher the semantic similarity between the Graph and the Text.

$$F_1 = 2\,\frac{p \cdot r}{p + r} \tag{1}$$

3.1 Training Data Creation

We aim to generate a training dataset of (graph, text, precision, recall) quadruples with a balanced and diverse distribution of precision and recall.

First we collect the set G of WebNLG graph/text pairs from the English WebNLG V3.0 dataset (Castro Ferreira et al., 2020)[5]. This dataset is semantically aligned (graph and text match in content), so we assign each $(g, t) \in G$ precision and recall scores of 1.

To increase precision and recall diversity we derive non aligned (g', t) pairs from $(g, t) \in G$ by pairing the text t with graphs g' which i) are subgraphs or super graphs of g or ii) modify g either by adding to it triples from non overlapping graphs or by modifying a triple contained in g. We then compute precision and recall for each new (g', t) pair based on the number of added, removed or modified triples.

Once we have a balanced English dataset, we extend it to other languages by machine translating the English texts. We use NLLB-200-3.3B (Team et al., 2022)[6] to translate into five languages: Irish, Maltese, Russian, Welsh, and Xhosa. To reduce the noise introduced by machine translation, we filter these translations following two criteria: Language Identification score via GlotLID (Kargaran et al., 2023)[7] and LID218e (Team et al., 2022)[8], and Semantic Similarity score via LaBSE (Feng et al., 2022)[9].

Figure 2 shows the distribution of precision and recall scores in 1.77 million (graph, text, precision, recall) quadruples evenly distributed across six languages. Appendix A provides more details.

[5] https://gitlab.com/shimorina/webnlg-dataset/-/tree/master/release_v3.0

[6] https://huggingface.co/facebook/nllb-200-3.3B

[7] https://huggingface.co/cis-lmu/glotlid

[8] https://huggingface.co/facebook/fasttext-language-identification

[9] https://huggingface.co/sentence-transformers/LaBSE

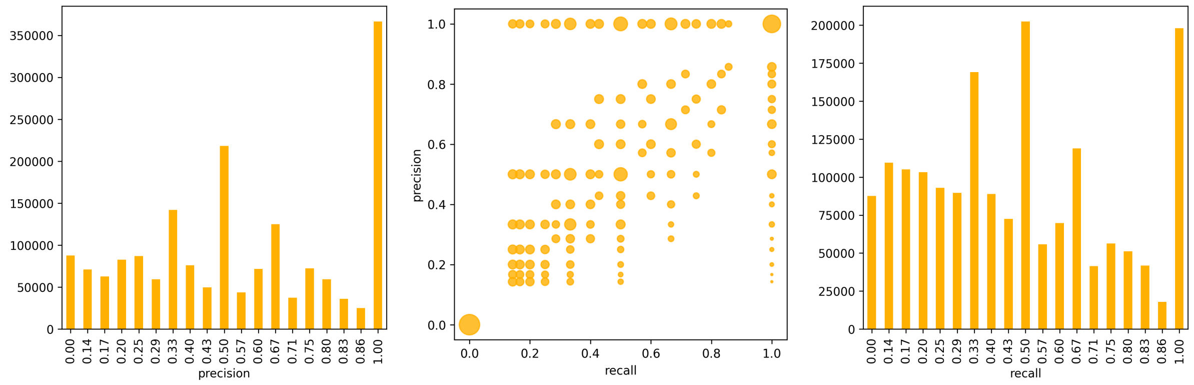


Figure 2: Number of samples by precision and recall scores in the training dataset.

3.2 Models

We use the following three models as baselines.

Data-QuestEval(DQE). A reference-less model by Rebuffel et al. (2021) which relies on question-generation and question-answering to assess semantic faithfulness. The main limitations of this baseline are that it was only fine-tuned for English, the long processing time of the text generation, and the risk of generating questions and answers unrelated to the actual input.

FactSpotter(FS). The latest model by Zhang et al. (2023)[10]. Compared to our approach, the main limitations of this baseline are that it was only fine-tuned for English and that it produces a single, recall oriented score.

NLI Base (NB). This baseline follows the exact same process as (Dušek and Kasner, 2020). We only change the off-the-shelf NLI model they used for a Multilingual one (Laurer et al., 2022)[11] instead of an English one. The main limitations of this baseline are that it is not familiar with the KG format and that it has not seen all the target languages we will test.

To learn our metric, we fine-tune mDeBERTa-v3-base-xnli-multilingual-nli-2mil7 (NB) on the dataset from Section 3.1. We fine-tune it as a regression model by targeting the entailment weights of the classification head and training simultaneously for precision and recall. For the precision score, we use the graph as premise and the text as hypothesis; for the recall score, we use the text as premise and the graph as the hypothesis. We fine-tune and compare three versions:

Multilingual Full Fine-Tuning (MultiFF): Full fine-tuning of the NLI Base model on all languages together.

Multilingual LoRA (MultiLR): LoRA on top of the NLI-Base model on all languages together.

Monolingual LoRA (MonoLR): Lora on top of the NLI-Base model for each language individually.

Appendix B provides details on the hyper parameters used to train our models.

4 Evaluation

We evaluate our metrics using correlation with human judgments (6 languages) and with automatic metrics (7 languages). We also report results on KG/text retrieval accuracy (7 languages). Table 2 summarises the test sets used.

4.1 Correlation with Automatic Metrics

Here, we use the 7L-Auto dataset, which consists of all graphs from the WebNLG test data[12] and all the texts generated from these graphs by participant systems of the WebNLG 2017, 2020, and 2023 Shared Tasks, as well as the different models trained by Meyer and Buys (2024). The models used to generate the texts include grammar-based and template-based approaches, statistical MT models, neural models trained from scratch, and fine-tuned pretrained models, covering a broad spectrum of errors and quality levels. Texts are generated in English, Russian, Breton, Irish, Maltese, Welsh, and Xhosa.

We compute the Spearman's Correlation (ρ) of the baselines and of our models with 5 reference-based metrics: BLEU (Papineni et al., 2002), ChrF++ (Popović, 2017), TER (Olive, 2005), BERTScore (Zhang* et al., 2020), and SBERT similarity (Reimers and Gurevych, 2019). For TER we report the inverse score ($\neg$TER $= 1 -$ TER) for easier display.

Dataset	# Graphs	# Texts	Languages	Relevant Annotations
7L-Auto	4461	143 838	br, cy, en, ga, mt, ru, xh	BLEU, ChrF++, TER, BERTScore, SBERT
4L-RP-Human	181	200	cy, en, mt, ru	Precision(p) and Recall(r)
2017	223	2 230	en	Semantics
2020	288	3 905	en, ru	Relevance(p), Correctness(p), Data Coverage(r)
2023	200	1 700	cy, ga, ru, mt	Omissions(r), Additions(r)

Table 2: Datasets used for Correlation Studies. When a relevant annotation is adjacent to Precision (p) or Recall (r) that is indicated, otherwise the annotation is consider adjacent to the more general F1 score.

[10] Inria-CEDAR/FactSpotter-DeBERTaV3-Base

[11] https://huggingface.co/MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7

[12] Specifically, we use the graphs from the WebNLG 2017 test set (1,862 graphs), from the WebNLG 2020 test set for English (1,779), from the WebNLG 2020 test set for Russian that are not present in the English test set (732) and from the Xhosa data sets that are not in any of the other datasets (88).

A good correlation would indicate that in the absence of ground truth, our metrics can be used as a proxy for these reference based metrics.

4.2 Correlation with Human Judgments

We use human judgments collected by the WebNLG campaigns and a subset we created to have human annotations specifically targeting precision and recall.

WebNLG 2017. The human annotations for this challenge (Shimorina et al., 2018) consist of 223 graphs lexicalized in English by 9 different NLG systems, plus the human-written references. The generations were scored on a 3-point Likert scale across 3 criteria: Fluency, Grammar, and Semantics. For this study, we focus on the Semantics annotation:

- Semantics: *Does the text correctly represent the meaning in the data?*

Since the Semantics annotation does not specify the type of error (Additions, Omission, etc.) we compute correlation between the WebNLG 2020 Semantics score and our F1 score.

WebNLG 2020. The human annotations for this challenge (Castro Ferreira et al., 2020) consist of 178 graphs lexicalized in English by 16 different NLG systems and 110 graphs lexicalized in Russian by seven different NLG systems; additionally, both include their human written references. The generations were scored on a 0 to 100 scale across five criteria: Text Structure, Fluency, Relevance, Correctness, and Data Coverage. For this study, we focus on the last three:

- Relevance: *Does the text describe only such predicates (with related subjects and objects), which are found in the data?*
- Correctness: *When describing predicates which are found in the data, does the text mention correct the objects and adequately introduces the subject for this specific predicate?*
- Data Coverage: *Does the text include descriptions of all predicates presented in the data?*

When computing correlation, we compare precision with the product of the Relevance and Correctness, recall with Data Coverage, and F1 with the harmonic means of both.

WebNLG 2023. The human annotations for this challenge (Cripwell et al., 2023) consist of 100 graphs lexicalized in Irish by 4 NLG systems, Maltese by 3 NLG systems, Welsh by 3 NLG systems, and other 100 graphs lexicalized in Russian by 3 NLG systems. Additionally, all of them included their human-written references. The generations were scored across 4 criteria: Fluency, Absence of Unnecessary Repetition, Absence of Additions, and Absence of Omissions. The first is on a 5-point Likert scale; the other 3 have binary Yes/No labels. For this study, we focus on Absence of Additions and Absence of Omissions:

- Absence of Additions: *Looking at the Text, is all of its content expressed in the Data expression? (Allow duplication of content.)*
- Absence of Omissions: *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)?*

We compute correlation between precision and Absence of Additions, recall and Absence of Omissions, and F1 with their harmonic mean.

4L-RP-Human. While WebNLG's existing human judgments can, to a certain extent, be used as proxies for Precision, Recall, and F1, none of them were collected to measure these values specifically. To address this, we created a new dataset of human judgments called 4L-RPHuman, with KG/text pairs extracted from the 7L-Auto dataset. It contains 50 KG/text pairs per language for four languages (English, Maltese, Russian, and Welsh) with a balanced distribution of Precision and Recall scores by our best-performing model. We then obtain human annotations for Precision and Recall of this subset to test how our model correlates with human judgments that specifically target these properties. The human annotators were provided with a text and a graph in table format and were asked to answer, using a scale of 1 to 5 (None, Few, Half, Most, All), the following questions:

- Precision: *How many Triples from the text can you find in the Table?*
- Recall: *How many Triples from the table can you find in the Text?*

The annotators were native speakers of the target language who were proficient in English hired via

Prolific[13] and paid 10£/h. Inter-annotator agreement was measured via Fleiss' Kappa (Fleiss, 1971). Appendix G provides more details.

4.3 Retrieval Accuracy

We also evaluate how well the scoring of various models discerns between good and bad pairings using a retrieval reformulation of the KG-to-text task: Given the embedding of a graph, how well can the model identify the most similar text in a corpus and vice versa given a text how well can it identify the corresponding graph?

Given a subset of 100 KG/text pairs randomly selected from the WebNLG dataset for each target language, we compute the F1 score with our model and the score produced by each of the baselines for each of the 10K possible graph/text combinations. We then compute Retrieval at 1 (A@1) i.e., the proportion of cases where the highest score is assigned to the correct graph-text pair. We limit the size of this subset given the computational demands of scoring all possible combinations of graphs and texts with our cross-encoder approach.

5 Results

5.1 Correlation with Automatic Metrics

Figure 3 shows the Spearman's Correlation (ρ) between the various reference-less metric we evaluate and reference-based automatic metrics on the 7L-Auto dataset (Breton, English, Irish, Maltese, Russian, Welsh and Xhosa WebNLG generations).

[13] https://www.prolific.com/

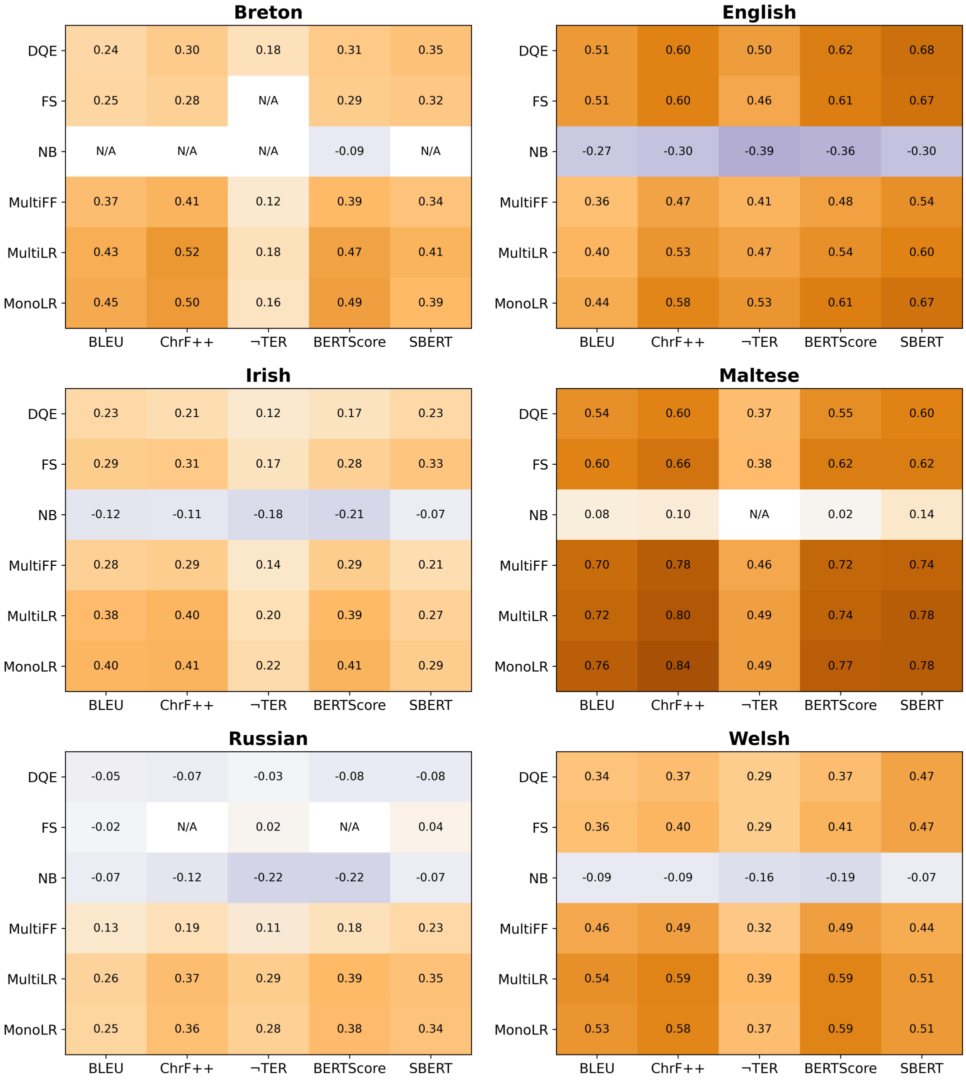


Figure 3: Spearman's Correlation (ρ) between reference-less and reference-based metrics on the 7L-Auto dataset. Only results with a p-value under 0.05 are reported. *NB: Since there is no Breton training data, the MonoLR score for Breton is computed with it's closest language (Welsh). Xhosa results are provided in* Appendix C.

Fine-tuning matters. Models that have been fine-tuned for the task (including the FS baseline) show positive correlation across metrics and languages, while the NB baseline has either negative or non significant correlations. These results highlight the limitations of using off-the-shelf models as proposed in (Dušek and Kasner, 2020) and underscore the importance of task specific fine-tuning.

Strong performance in English. While trained on multilingual data, our models almost match the performance of metrics trained on English only (DQE, FS, NB). Interestingly, the gap is smallest for semantic based metrics (SBERT, BERTScore), suggesting that our metrics are good at capturing paraphrases.

Good performance in other languages. Our fine-tuned models, especially the small Monolingual LoRA version, outperform all three baselines in all the other languages. These results demonstrate the effectiveness of our approach despite fine-tuning on synthetic, non-gold data.

5.2 Correlation with Human Judgments

5.2.1 Evaluation on Human Judgments for Precision and Recall.

Table 3 reports correlation results when comparing the precision, recall and F1 scores predicted by our models with corresponding human judgments (4L-RP-Human dataset) and the inter-annotator agreement (Fleiss κ). They show a strong correlation for all three metrics in English, Russian and Welsh and a moderate one for Maltese showing the effectiveness of our approach to capture omissions (recall), addition (precision) and semantic faithfulness (F_1).

We provide examples illustrating good, medium and bad output from our best model (MonoLR) for each target language in Table 13, Table 14, Table 15, and Table 16 (Appendix F).

5.2.2 Evaluation on Human Judgments from the WebNLG Shared Tasks.

Figure 4 shows the Root Mean Squared Error (RMSE) and Spearman's correlation (ρ) of the F1 score of different automatic metrics against the WebNLG 2017, 2020 and 2023 human annotations.

Language	**Annotators**	**Precision** Fleiss κ	ρ	**Recall** Fleiss κ	ρ	**F1** ρ
English	4	0.47	0.68	0.47	0.63	0.70
Maltese	3	0.29	0.38	0.49	0.30	0.47
Russian	2	0.32	0.63	0.39	0.52	0.67
Welsh	4	0.37	0.60	0.50	0.81	0.70

Table 3: Fleiss' κ of Precision and Recall Human Judgments as well as the Spearman's Correlation (ρ) of their average compared to our MonoLR model on the 4L-RP-Human subset

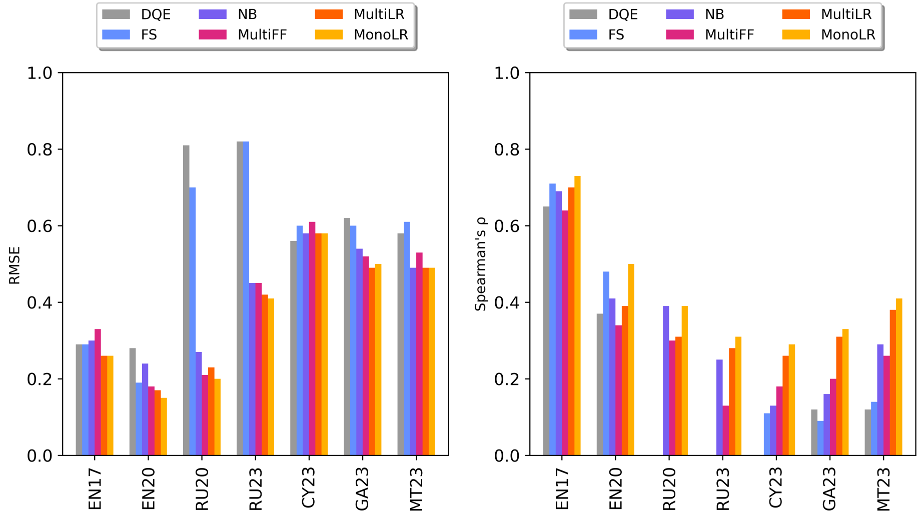


Figure 4: Root Mean Squared Error (RMSE) and Spearman's correlation (ρ) of the F1 score from different automatic metrics against the closest approximate human annotations from *WebNLG 2017, 2020 and 2023* (Each year has different annotations, see Table 2 for more details). For the Spearman's correlation scores, only results with a p-value under 0.05 are reported.

	A@1						
	Breton	English	Irish	Maltese	Russian	Welsh	Xhosa
DQE	0.53	0.95	0.32	0.48	0.05	0.39	0.38
FS	0.37	0.99	0.36	0.45	0.11	0.46	0.24
NB	0.56	0.79	0.49	0.60	0.70	0.60	0.68
MultiFF	0.92	**1.00**	0.84	**0.96**	0.96	0.95	**0.99**
MultiLR	0.93	0.99	0.85	0.94	0.93	0.97	0.98
MonoLR	**0.94**	**1.00**	**0.91**	**0.96**	**0.99**	**1.00**	**0.99**

Table 4: Retrieval at 1 (A@1) when using the F1 score to match graphs with their corresponding text on 100 selected examples from the 7L-Auto dataset. *NB: Since there is no Breton training data, the MonoLR score for Breton is computed with it's closest language (Welsh).*

The exact numbers as well as a breakdown by Precision and Recall (when possible), can be found in Table 7, Table 8, and Table 9 (Appendix D).

Correlation is highest for 2017 Data. Different from our 4L-RP-Human dataset, the human judgments collected during the WebNLG campaigns do not directly target precision and recall: WebNLG 2017 targets semantic faithfulness, WebNLG 2020 targets three semantic criteria related to but not identical to precision and recall, and WebNLG 2023 focuses on omissions and additions but only return a binary score no matter how much omission/addition occurs in the generated text. As explained in Section 4.2, we use the available human scores to approximate an F1 score and compute correlation between these derived F1 scores and each evaluated metric. We hypothesize that the higher correlation obtained for the 2017 data results from the fact that for this campaign, the single score provided by the human evaluation is more directly related to the unique score provided by the baseline metrics and to our F1 score. Conversely the lower correlation scores obtained by our models on the WebNLG 2020 and 2023 datasets compared to those obtained when evaluating on the 4L-RP-Human dataset are likely caused by the need to "reconstruct" an F1 score from the human judgments provided by these datasets (product of three criteria for 2020 and Harmonic mean of binary scores for lack of addition and omission for 2023).

An improvement over the state-of-the-art. In English, our MonoLR model outperforms the three baselines despite these being optimized for this language. For the other 4 languages, the gap with these monolingual metrics is particularly pronounced. Surprisingly, for Russian the NB model is on par with our MonoLR model. This highlights the impact of using a multilingual model as base model even when fine tuning on English only. However, the low results of the NB model on the other languages shows that using NLI only, without fine tuning on task specific data does not suffice.

5.3 Accuracy on Retrieval.

Table 4 shows the Retrieval at 1 (A@1) for text/graph retrieval. In all languages, our fine-tuned models outperform the baselines, with the MonoLR model obtaining almost perfect scores in most of them and even outperforming FS in English. While the retrieval corpus is admittedly limited in size (10K possible combinations), the results demonstrate the effect of our approach on multilingual graph/text representation learning: for all languages, our models successfully identify the matching text given a graph and vice versa.

6 Conclusion

Previous work on reference less evaluation of KG-to-Text generation has mainly focused on English, providing global metrics for semantic faithfulness. We extend this work by presenting models which support the reference less evaluation of multilingual KG-to-Text generation while allowing for a finer-grained evaluation in terms of precision, recall and F1. The proposed models show strong correlation with human judgments of precision and recall for several languages, moderate to strong correlation with automatic metrics and high retrieval accuracy. On a small data set of 10K (graph,text) pairs retrieval accuracy is high indicating that the representations learned by our model provide a good basis for identifying matching KG/text pairs.

7 Acknowledgments

We thank the anonymous reviewers for their feedback. We gratefully acknowledge the support of the French National Research Agency (Gardent; award ANR-20-CHIA-0003, XNLG "Multilingual, Multi-Source Text Generation"). Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000.fr). This work was also granted access to the HPC resources of IDRIS under the allocation AD0110165612 made by GENCI.

8 Limitations

First and foremost we were limited by the availability of high quality English KG-to-Text data. While we could generate infinite synthetic errors our original source of correct English lexicalizations was short and could not be securely extended.

We were limited by available multilingual data. While machine translation has advanced significantly over time and we attempted to filter out bad translations, there is still a risk of noise getting into the training data; particularly on low-resource languages. Furthermore, the lack of adequate testing data means that, while we believe it is possible to apply this framework to many more languages, we were unable to put it to the test.

Our trained models are limited to the languages we trained them for and are still prone to mistake in certain cases, particularly when dealing with the grammatical and morphological nuances of languages.

9 Ethical Considerations

Expanding the toolbox of evaluation metrics to languages other than English can help democratize and expand access to new technologies to a larger and more diverse group of people; however, it is important to keep in mind the limitations of these systems. Neural Models are, at the end of the day, statistical models and as such they are prone to error.

It is important to always consider the multiple layers of bias and noise involved in the creation of these models. Every step of the way is a possible source of bias; from the selected pretrained model to the fine-tuning dataset, going through all the intermediary processes like synthetic data generation and machine translations.

In a time of quick changes it is more important than ever to have multiple and diverse evaluation metrics that provide different insights into the generation of our models and remember that each and everyone of those metrics has its own pros and cons. We can not blindly trust a single specific metric but instead understand and use a multitude of them to gain a complete perspective.

Supplementary Materials Availability: All the code, data, and final models produced are made publicly available in the following repository: https://gitlab.inria.fr/wsotomar/semantic-evaluation-of-multilingual-data-to-text-generation-via-nli-fine-tuning.

References

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.

Yanran Chen and Steffen Eger. 2023. MENLI: Robust evaluation metrics from natural language inference. *Transactions of the Association for Computational Linguistics*, 11:804–825.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

Liam Cripwell, Anya Belz, Claire Gardent, Albert Gatt, Claudia Borg, Marthese Borg, John Judge, Michela Lorandi, Anna Nikiforovskaya, and William Soto Martinez. 2023. The 2023 WebNLG shared task on low resource languages. overview and evaluation results (WebNLG 2023). In *Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023)*, pages 55–66, Prague, Czech Republic. Association for Computational Linguistics.

Ondřej Dušek and Zdeněk Kasner. 2020. Evaluating semantic accuracy of data-to-text generation with natural language inference. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 131–137, Dublin, Ireland. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

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.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Hassan Kane, Muhammed Yusuf Kocyigit, Ali Abdalla, Pelkins Ajanoh, and Mohamed Coulibali. 2020. NUBIA: NeUral based interchangeability assessor for text generation. In *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*, pages 28–37, Online (Dublin, Ireland). Association for Computational Linguistics.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. GlotLID: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218, Singapore. Association for Computational Linguistics.

Moritz Laurer, Wouter van Atteveldt, Andreu Salleras Casas, and Kasper Welbers. 2022. Less Annotating, More Classifying – Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT - NLI. *Preprint*. Publisher: Open Science Framework.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Francois Meyer and Jan Buys. 2024. Triples-to-isiXhosa (T2X): Addressing the challenges of low-resource agglutinative data-to-text generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16841–16854, Torino, Italia. ELRA and ICCL.

Anna Nikiforovskaya and Claire Gardent. 2024. Evaluating RDF-to-text generation models for English and Russian on out of domain data. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 134–144, Tokyo, Japan. Association for Computational Linguistics.

Joseph Olive. 2005. Global autonomous language exploitation (gale). *DARPA/IPTO Proposer Information Pamphlet*.

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.

Clement Rebuffel, Thomas Scialom, Laure Soulier, Benjamin Piwowarski, Sylvain Lamprier, Jacopo Staiano, Geoffrey Scoutheeten, and Patrick Gallinari. 2021. Data-QuestEval: A referenceless metric for data-to-text semantic evaluation. In *Proceedings of* *the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8029–8036, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Anastasia Shimorina, Claire Gardent, Shashi Narayan, and Laura Perez-Beltrachini. 2018. WebNLG Challenge: Human Evaluation Results. Technical report, Loria & Inria Grand Est.

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. *Preprint*, arXiv:2207.04672.

Huajian Zhang and Laura Perez-Beltrachini. 2024. Leveraging entailment judgements in cross-lingual summarisation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14481–14497, Bangkok, Thailand. Association for Computational Linguistics.

Kun Zhang, Oana Balalau, and Ioana Manolescu. 2023. FactSpotter: Evaluating the factual faithfulness of graph-to-text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10025–10042, Singapore. Association for Computational Linguistics.

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

A Training Data Creation: Expanded

A.1 Quadruple Collection

As stated before, we begin our data creation process with the English WebNLG V3.0 dataset (Castro Ferreira et al., 2020)[14]. This dataset contains aligned (graph g_i, text t_i) pairs. We can also refer to graph g_i as g_{t_i}, meaning it is the graph aligned with t_i. In the dataset, the graphs were extracted from DBPedia,[15] and the texts were either automatically lexicalized or mined from Wikipedia[16] before being aligned with each other by human annotators. Since graph and text are aligned the pair has a precision and recall scores of 1, forming a quadruple (graph g_{t_i}, text t_i, precision=1, recall=1).

We can create variations of these original quadruples with diverse precision and recall scores by finding pairs (graph g_j, text t_i) with different levels of information overlap (o). To do so, we propose to keep the text static and change the graph, since it is much easier to work with and manipulate data in graph representation. For example, measuring o is much easier when both elements are in graph representation, since we can just compute the intersection between both sets of triples ($o = |g_j \cap g_{t_i}|$). Because of that, for most of the creation process we work with (graph g_j, graph g_{t_i}) pairs instead of (graph g_j, text t_i) pairs. Only at the end of the process we substitute g_{t_i} back with the original text t_i.

Starting with our graph g_{t_i}, if we wish to obtain a variation quadruples with precision p and recall r, we need to find a new g_j such that the following equations are true:

- $o/|g_{t_i}| = p$
- $o/|g_j| = r$

At first, we look for such a g_j in the list of all original graphs from WebNLGand all its subgraphs. If finding a matching graph is impossible, we create a synthetic one that satisfies the criteria.

A.2 Synthetic Graph Creation

Given a graph g_{t_i}, we can create a synthetic graph g_j with precision p and recall r by first taking o triples from g_{t_i} and then adding external triples to g_j until $o/|g_j| = r$.

External triples can be procured by selecting a triple from some graph g_k that does no overlap with g_{t_i} or by corrupting real triples from g_{t_i} so that the information they represent does not match the original graph.

When corrupting a real triple, we can swap the order of the elements in the triple or substitute some or all of its elements with incorrect values. When doing so, we use logical substitutions. For example, to corrupt the triple (Alan Bean | birthPlace | Wheeler, Texas), we can substitute the object *Wheeler, Texas* with a different value. In such a case, we would select a value that can be paired with the property *birthPlace*, like *Miami, Florida*, instead of a random value like *1932-03-15*.

A.3 Toy Example

If we start with the following dataset of aligned (graph g_i, text t_i) pairs:

ID	Graph	Text
1	Alice | occupation | Writer	Alice is a writer.
2	Alice | occupation | Writer Alice | country | USA	Alice is an American writer.
3	Alice | country | USA Bob | country | USA	Alice and Bob are Americans.

We can assign all of them precision and recall values of 1 and turn them into quadruples.

ID	Graph	Text	P	R
1	Alice | occupation | Writer	Alice is a writer.	1.00	1.00
2	Alice | occupation | Writer Alice | country | USA	Alice is an American writer.	1.00	1.00
3	Alice | country | USA Bob | country | USA	Alice and Bob are Americans.	1.00	1.00

To create new quadruples we can start by pairing texts with subgraphs or supergraphs of their original graph. For example, g_1 is a subgraph of g_2 (and therefore g_2 is a supergraph of g_1). Pairing a text with a supergraph will produce a quadruple where the text is missing information (omission), leading to a lower recall. Paring a text with a subgraph will produce a quadruple where the text has extra information (addition/hallucination), leading to a lower precision:

ID	Graph	Text	P	R
1	Alice | occupation | Writer	Alice is a writer.	1.00	1.00
2	Alice | occupation | Writer Alice | country | USA	Alice is an American writer.	1.00	1.00
3	Alice | country | USA Bob | country | USA	Alice and Bob are Americans.	1.00	1.00
4	Alice | occupation | Writer	Alice is an **American** writer.	0.50	1.00
5	Alice | occupation | Writer **Alice | country | USA**	Alice is a writer.	1.00	0.50

We can also pair a text with partially overlapping graphs. For example, g_2 and g_3 overlap in

[14] https://gitlab.com/shimorina/webnlg-dataset/-/tree/master/release_v3.0

[15] https://www.dbpedia.org/

[16] https://www.wikipedia.org/

one triple, by matching their texts and graphs we will have quadruples where both recall and precision can be affected (there are both omissions and additions/hallucinations):

ID	Graph	Text	P	R
1	Alice | occupation | Writer	Alice is a writer.	1.00	1.00
2	Alice | occupation | Writer Alice | country | USA	Alice is an American writer.	1.00	1.00
3	Alice | country | USA Bob | country | USA	Alice and Bob are Americans.	1.00	1.00
4	Alice | occupation | Writer	Alice is an **American** writer.	0.50	1.00
5	Alice | occupation | Writer **Alice | country | USA**	Alice is a writer.	1.00	0.50
6	Alice | country | USA **Bob | country | USA**	Alice is an American **writer.**	0.50	0.50
7	**Alice | occupation | Writer** Alice | country | USA	Alice **and Bob** are Americans.	0.50	0.50

Finally, we can produce new quadruples by creating synthetic graphs, either by corrupting original triples or by adding new ones:

ID	Graph	Text	P	R
1	Alice | occupation | Writer	Alice is a writer.	1.00	1.00
2	Alice | occupation | Writer Alice | country | USA	Alice is an American writer.	1.00	1.00
3	Alice | country | USA Bob | country | USA	Alice and Bob are Americans.	1.00	1.00
4	Alice | occupation | Writer	Alice is an **American** writer.	0.50	1.00
5	Alice | occupation | Writer **Alice | country | USA**	Alice is a writer.	1.00	0.50
6	Alice | country | USA **Bob | country | USA**	Alice is an American **writer.**	0.50	0.50
7	**Alice | occupation | Writer** Alice | country | USA	Alice **and Bob** are Americans.	0.50	0.50
8	Alice | occupation | Writer Alice | country | **Mexico**	Alice is an **American** writer.	0.50	0.50
9	Alice | occupation | Writer Alice | country | USA **Alice | birthDate | 2000-01-01**	Alice is an American writer.	1.00	0.66

B Training Hyper Parameters

Parameter	**MultiFF***	**MultiLR***	**MonoLR****
Training Hardware	1 32GB V100	1 32GB V100	1 32GB V100
Training Instances	$\sim$ 3 544 994	$\sim$ 3 544 994	$\sim$ 590 832
Training Epochs	1	1	1
Training Time	$\sim$ 7h	$\sim$ 11h	$\sim$ 2h
Warmup Steps	10%	10%	10%
Scheduler	WarmupLinear	WarmupLinear	WarmupLinear
Optimizer	AdamW	AdamW	AdamW
Learning Rate	2e-5	2e-5	2e-5
Loss Function	BCELoss	BCELoss	BCELoss
Rank	N/A	N/A	32
Total parameters	278 811 651	283 507 299	283 507 299
Trained parameters	278 811 651	5 382 240	5 382 240

Table 5: Training hyperparameters of all our models. *The model is fine-tuned on 6 languages together. **The model is fine-tuned in only 1 language.

C 7L-Auto Results

	Breton					**English**				
			F1					F1		
	BLEU ↑	ChrF++ ↑	¬TER ↑	BERTScore ↑	SBERT ↑	BLEU ↑	ChrF++ ↑	¬TER ↑	BERTScore ↑	SBERT ↑
DQE	0.24	0.30	**0.18**	0.31	0.35	**0.51**	**0.60**	0.50	**0.62**	**0.68**
FS	0.25	0.28	—	0.29	0.32	**0.51**	**0.60**	0.46	0.61	0.67
NB	—	—	—	-0.09	—	-0.27	-0.30	-0.39	-0.36	-0.30
MultiFF	0.37	0.41	0.12	0.39	0.34	0.36	0.47	0.41	0.48	0.54
MultiLR	0.43	**0.52**	**0.18**	0.47	**0.41**	0.40	0.53	0.47	0.54	0.60
MonoLR	**0.45**	0.50	0.16	**0.49**	0.39	0.44	0.58	**0.53**	0.61	0.67

	Irish					**Maltese**				
			F1					F1		
	BLEU ↑	ChrF++ ↑	¬TER ↑	BERTScore ↑	SBERT ↑	BLEU ↑	ChrF++ ↑	¬TER ↑	BERTScore ↑	SBERT ↑
DQE	0.23	0.21	0.12	0.17	0.23	0.54	0.60	0.37	0.55	0.60
FS	0.29	0.31	0.17	0.28	**0.33**	0.60	0.66	0.38	0.62	0.62
NB	-0.12	-0.11	-0.18	-0.21	-0.07	0.08	0.10	—	0.02	0.14
MultiFF	0.28	0.29	0.14	0.29	0.21	0.70	0.78	0.46	0.72	0.74
MultiLR	0.38	0.40	0.20	0.39	0.27	0.72	0.80	**0.49**	0.74	**0.78**
MonoLR	**0.40**	**0.41**	**0.22**	**0.41**	0.29	**0.76**	**0.84**	**0.49**	**0.77**	**0.78**

	Russian					**Welsh**				
			F1					F1		
	BLEU ↑	ChrF++ ↑	¬TER ↑	BERTScore ↑	SBERT ↑	BLEU ↑	ChrF++ ↑	¬TER ↑	BERTScore ↑	SBERT ↑
DQE	-0.05	-0.07	-0.03	-0.08	-0.08	0.34	0.37	0.29	0.37	0.47
FS	-0.02	—	0.02	—	0.04	0.36	0.40	0.29	0.41	0.47
NB	-0.07	-0.12	-0.22	-0.22	-0.07	-0.09	-0.09	-0.16	-0.19	-0.07
MultiFF	0.13	0.19	0.11	0.18	0.23	0.46	0.49	0.32	0.49	0.44
MultiLR	**0.26**	**0.37**	**0.29**	**0.39**	**0.35**	**0.54**	**0.59**	**0.39**	**0.59**	**0.51**
MonoLR	0.25	0.36	0.28	0.38	0.34	0.53	0.58	0.37	**0.59**	**0.51**

	Xhosa				
			F1		
	BLEU ↑	ChrF++ ↑	¬TER ↑	BERTScore ↑	SBERT ↑
DQE	-0.10	-0.04	-0.14	-0.11	-0.12
FS	0.19	0.18	0.18	0.11	0.13
NB	-0.25	-0.27	-0.30	-0.24	-0.26
MultiFF	—	-0.05	-0.11	—	-0.05
MultiLR	0.19	0.32	0.15	0.22	0.21
MonoLR	**0.22**	**0.34**	**0.19**	**0.26**	**0.25**

Table 6: Spearman's Correlation (ρ) of the F1 score from different automatic metrics against classic reference-based metrics on the 7L-Auto dataset. Only results with a p-value under 0.05 are reported. *NB: Since there is no Breton training data, the MonoLR score for Breton is computed with it's closest language (Welsh).*

D WebNLG Human Judgment Results

	English	
	F1	
	RMSE ↓	ρ ↑
DQE	0.29	0.65
FS	0.29	0.71
NB	0.30	0.69
MultiFF	0.33	0.64
MutiLR	**0.26**	0.70
MonoLR	**0.26**	**0.73**

Table 7: Root Mean Squared Error (RMSE) and Spearman's correlation (ρ) of the F1 score from different automatic metrics against the *English WebNLG 2017* human annotations. For the Spearman's correlation scores, only results with a p-value under 0.05 are reported.

	English						Russian					
	P		R		F1		P		R		F1	
	RMSE ↓	ρ ↑	RMSE ↓	ρ ↑	RMSE ↓	ρ ↑	RMSE ↓	ρ ↑	RMSE ↓	ρ ↑	RMSE ↓	ρ ↑
DQE	0.27	0.37	0.32	0.32	0.28	0.37	0.79	—	0.84	—	0.81	—
FS	0.21	0.35	0.18	0.45	0.19	0.48	0.67	—	0.79	—	0.70	—
NB	0.28	0.28	0.16	0.43	0.24	0.41	0.30	0.21	0.20	0.39	0.27	0.39
MultiFF	0.22	0.22	0.15	0.38	0.18	0.34	0.25	0.14	**0.14**	0.42	0.21	0.30
MultiLR	0.22	0.36	0.20	0.37	0.17	0.39	0.26	0.19	0.22	0.36	0.23	0.31
MonoLR	**0.20**	**0.44**	**0.14**	**0.47**	**0.15**	**0.50**	**0.24**	**0.25**	0.16	**0.44**	**0.20**	**0.39**

Table 8: Root Mean Squared Error (RMSE) and Spearman's correlation (ρ) of the Precision, Recall, and F1 score from different automatic metrics against the *English and Russian WebNLG 2020* human annotations. For the Spearman's correlation scores, only results with a p-value under 0.05 are reported.

	Irish						Maltese					
	P		R		F1		P		R		F1	
	RMSE ↓	ρ ↑	RMSE ↓	ρ ↑	RMSE ↓	ρ ↑	RMSE ↓	ρ ↑	RMSE ↓	ρ ↑	RMSE ↓	ρ ↑
DQE	0.65	0.11	0.64	0.09	0.62	0.12	0.60	0.14	0.59	0.10	0.58	0.12
FS	0.62	—	0.57	0.13	0.60	0.09	0.66	0.13	0.52	0.17	0.61	0.14
NB	0.52	0.14	0.46	0.22	0.54	0.16	0.48	0.20	0.47	0.37	0.49	0.29
MultiFF	0.48	0.14	0.49	0.29	0.52	0.20	0.46	0.14	0.52	0.32	0.53	0.26
MultiLR	0.47	0.18	**0.45**	0.35	**0.49**	0.31	0.44	0.26	0.46	0.37	**0.49**	0.38
MonoLR	**0.46**	**0.21**	**0.45**	**0.37**	0.50	**0.33**	**0.43**	**0.30**	**0.45**	**0.43**	**0.49**	**0.41**

	Russian						Welsh					
	P		R		F1		P		R		F1	
	RMSE ↓	ρ ↑	RMSE ↓	ρ ↑	RMSE ↓	ρ ↑	RMSE ↓	ρ ↑	RMSE ↓	ρ ↑	RMSE ↓	ρ ↑
DQE	0.85	—	0.87	—	0.82	—	0.61	—	0.58	0.11	0.56	—
FS	0.76	—	0.86	—	0.82	—	0.69	—	0.55	0.18	0.60	0.11
NB	0.45	0.17	0.34	0.28	0.45	0.25	0.54	0.13	**0.53**	0.24	**0.58**	0.13
MultiFF	0.42	—	0.36	0.22	0.45	0.13	0.51	—	0.58	0.26	0.61	0.18
MultiLR	0.39	0.24	0.35	0.28	0.42	0.28	**0.49**	0.18	0.54	0.29	**0.58**	0.26
MonoLR	**0.38**	**0.25**	**0.32**	**0.33**	**0.41**	**0.31**	**0.49**	**0.21**	0.54	**0.34**	**0.58**	**0.29**

Table 9: Root Mean Squared Error (RMSE) and Spearman's correlation (ρ) of the Precision, Recall, and F1 score from different automatic metrics against the *Irish, Maltese, Russian and Welsh WebNLG 2023* human annotations. For the Spearman's correlation scores, only results with a p-value under 0.05 are reported. *NB: The original human annotations are binary labels, which might explain the high RMSE*

E WebNLG Human Judgment Results: System-level Correlations

For comparison, we also report results obtained using the script from Zhang et al. (2023) for system level correlations.

	English		
	Semantic		
	$r\uparrow$	$\rho\uparrow$	$\tau\uparrow$
*OgFS**	**0.97**	**0.93**	**0.85**
FS	0.95	0.91	0.80
NB	0.96	0.91	0.80
MultiFF	0.90	0.82	0.66
MultiLR	0.93	0.84	0.70
MonoLR	0.96	0.92	0.81

Table 10: System-level correlation (Pearson's r, Spearman's ρ, and Kendall's τ) of the F1 score from different automatic metrics against the *English WebNLG 2017* human annotations. *OgFS are the results reported in the original FactSpotter paper.

	English								
	Correctness			Relevance			Data Coverage		
	$r\uparrow$	$\rho\uparrow$	$\tau\uparrow$	$r\uparrow$	$\rho\uparrow$	$\tau\uparrow$	$r\uparrow$	$\rho\uparrow$	$\tau\uparrow$
*OgFS**	0.94	0.80	0.64	0.96	0.79	0.64	0.91	0.87	0.71
FS	**0.97**	**0.84**	**0.70**	**0.98**	**0.80**	**0.65**	**0.97**	**0.94**	**0.83**
NB	0.94	0.75	0.60	0.91	0.69	0.53	0.95	0.90	0.77
MultiFF	0.93	0.75	0.60	0.95	0.77	0.62	0.93	0.92	0.79
MultiLR	0.92	0.76	0.60	0.92	0.78	0.62	0.92	0.89	0.74
MonoLR	0.93	0.75	0.60	0.94	0.78	0.64	0.94	**0.94**	**0.83**

	Russian								
	Correctness			Relevance			Data Coverage		
	$r\uparrow$	$\rho\uparrow$	$\tau\uparrow$	$r\uparrow$	$\rho\uparrow$	$\tau\uparrow$	$r\uparrow$	$\rho\uparrow$	$\tau\uparrow$
FS	-0.82	0.79	—	0.62	0.63	**0.81**	—	0.83	0.75
NB	**0.80**	**0.83**	**0.75**	**0.83**	**0.85**	0.76	0.83	0.85	0.76
MultiFF	—	—	—	0.79	0.82	0.74	**0.95**	0.88	0.78
MultiLR	—	—	—	0.80	0.83	0.74	0.94	**0.89**	0.79
MonoLR	—	—	—	0.80	0.83	0.76	0.92	**0.89**	**0.80**

Table 11: System-level correlation (Pearson's r, Spearman's ρ, and Kendall's τ) of the Precision and Recall score from different automatic metrics against the *English and Russian WebNLG 2020* human annotations. *OgFS are the results reported in the original FactSpotter paper.

	Irish						Maltese					
	Additions			Omissions			Additions			Omissions		
	$r\uparrow$	$\rho\uparrow$	$\tau\uparrow$	$r\uparrow$	$\rho\uparrow$	$\tau\uparrow$	$r\uparrow$	$\rho\uparrow$	$\tau\uparrow$	$r\uparrow$	$\rho\uparrow$	$\tau\uparrow$
FS	0.79	0.60	—	0.97	0.97	—	0.82	0.79	—	**1.00**	**1.00**	—
NB	0.97	**1.00**	—	0.97	**1.00**	—	**1.00**	0.99	—	0.97	0.99	—
MultiFF	0.97	**1.00**	—	0.97	**1.00**	—	**1.00**	**1.00**	—	**1.00**	**1.00**	—
MultiLR	0.97	**1.00**	—	0.96	**1.00**	—	**1.00**	**1.00**	—	**1.00**	**1.00**	—
MonoLR	**0.98**	**1.00**	—	**0.98**	**1.00**	—	**1.00**	**1.00**	—	**1.00**	**1.00**	—

	Russian						Welsh					
	Additions			Omissions			Additions			Omissions		
	$r\uparrow$	$\rho\uparrow$	$\tau\uparrow$	$r\uparrow$	$\rho\uparrow$	$\tau\uparrow$	$r\uparrow$	$\rho\uparrow$	$\tau\uparrow$	$r\uparrow$	$\rho\uparrow$	$\tau\uparrow$
FS	-0.48	-0.43	—	0.61	0.47	—	**0.97**	**0.84**	—	**1.00**	**1.00**	—
NB	**0.25**	**0.15**	—	**0.90**	**0.71**	—	**0.97**	**0.84**	—	**1.00**	0.96	—
MultiFF	-0.66	-0.49	—	0.68	0.64	—	**0.97**	**0.84**	—	**1.00**	**1.00**	—
MultiLR	-0.80	-0.60	—	-0.41	-0.11	—	**0.97**	**0.84**	—	**1.00**	**1.00**	—
MonoLR	-0.48	0.05	—	0.45	0.48	—	**0.97**	**0.84**	—	**1.00**	**1.00**	—

Table 12: System-level correlation (Pearson's r, Spearman's ρ, and Kendall's τ) of the Precision and Recall score from different automatic metrics against the *Irish, Maltese, Russian and Welsh WebNLG 2023* human annotations.

F Annotation Examples

Sample	Precision		Recall		Quintile
	Human	MonoLR	Human	MonoLR	
Graph: Mermaid (Train song) | genre | Pop rock Mermaid (Train song) | runtime | 3.16 Mermaid (Train song) | releaseDate | 2012-12-27 Mermaid (Train song) | precededBy | This'll Be My Year Mermaid (Train song) | writer | Espen Lind **Text:** Mermaid is a pop rock song written by Espen Lind. It was released on 27 December 2012 and has a run time of 3.16.	1.00	0.98	0.62	0.75	1st
Graph: Turkey | longName | Republic of Turkey Nurhan Atasoy | nationality | Turkish people Nurhan Atasoy | citizenship | Turkey Turkey | language | Turkish language **Text:** The Turkish language is spoken in Turkey where the leader is known as the Republic of Turkey. The country is the location of the Ataturk Atasoy which is a citizenship of the Turkish people.	0.19	0.32	0.25	0.37	3rd
Graph: Ciudad Ayala | populationMetro | 1777539 **Text:** 1777539 is the population metro in the country.	0.12	0.88	0.25	0.52	5th

Table 13: English samples from 4L-RP-Human with their Human and MonoLR scores in a scale from 0 to 1. The samples were selected from the 1st (good), 3rd (median) and 5th (bad) quintile based on the accuracy of MonoLR compared to the Human judgment.

Sample	Precision		Recall		Quintile
	Human	MonoLR	Human	MonoLR	
Graph: McVeagh of the South Seas | director | Cyril Bruce McVeagh of the South Seas | writer | Harry Carey (actor born 1878) **Text:** McVeagh tal-Baħar tan-Nofsinhar kien miktub minn Harry Carey (attur imwieled 1878) u dirett minn Cyril Bruce. **English MT:** McVeagh of the Southern Seaboard was written by Harry Carey (actor born 1878) and directed by Cyril Bruce.	0.92	0.96	0.83	0.88	1st
Graph: United States | leaderTitle | Vice President Darinka Dentcheva | residence | United States **Text:** Il-Viċi President huwa l-mexxej tal-Istati Uniti **English MT:** The Vice President is the leader of the United States.	0.75	0.99	0.42	0.67	3rd
Graph: Turkey | demonym | Turk **Text:** Id-demonimu tal-abitanti tal-belt ta' Turkmen huwa ta **English MT:** The demonym of the inhabitants of the city of Turkmen is	0.08	0.84	0.08	0.71	5th

Table 14: Maltese samples from 4L-RP-Human with their Human and MonoLR scores in a scale from 0 to 1. The samples were selected from the 1st (good), 3rd (median) and 5th (bad) quintile based on the accuracy of MonoLR compared to the Human judgment.

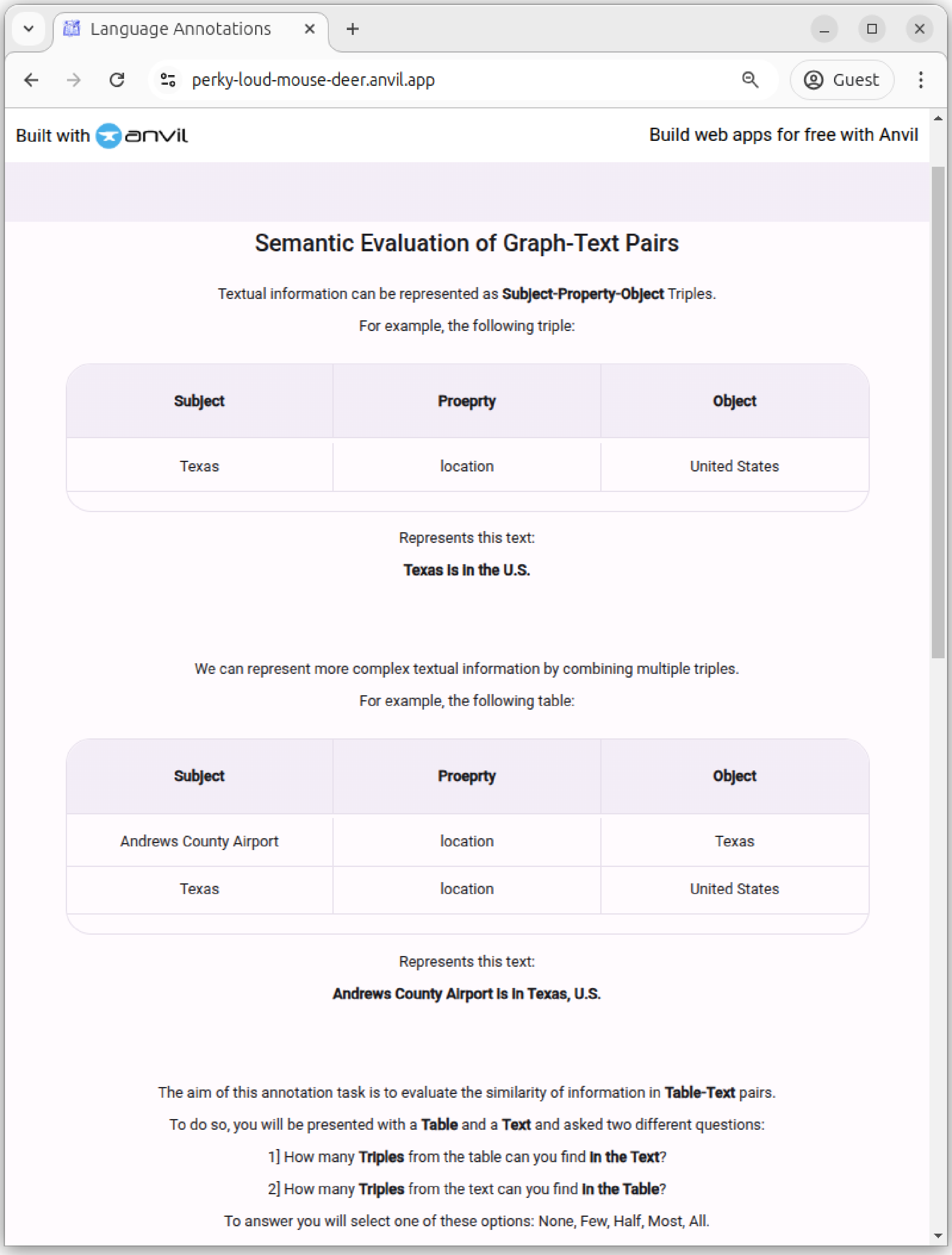
Sample	Precision		Recall		Quintile
	Human	MonoLR	Human	MonoLR	
Graph: (66063) 1998 RO1 | meanTemperature | 265.0 (kelvins) (66063) 1998 RO1 | apoapsis | 254989570.60815 (kilometres) (66063) 1998 RO1 | epoch | 2013-11-04 (66063) 1998 RO1 | orbitalPeriod | 360.29 (days) **Text:** Небесное тело, известное как (66063) 1998 RO1, имеет среднюю температуру 265 Кельвинов и орбитальный период 360,29 дней. Его апоцентр - 254989570,60815 километров, а его эпоха - 13 января 2016 года. **English MT:** The celestial body, known as (66063) 1998 RO1, has an average temperature of 265 Kelvin and an orbital period of 360.29 days. Its apocenter is 254989570.60815 kilometers, and its epoch is January 13, 2016.	0.88	0.93	0.88	0.79	1st
Graph: School of Business and Social Sciences at the Aarhus University | affiliation | European University Association European University Association | headquarter | Brussels School of Business and Social Sciences at the Aarhus University | established | 1928 **Text:** Школа бизнеса и социальных наук Орхусского университета была создана в 1928 году и входит в Ассоциацию университетов Европы, штаб-квартира которой находится в Брюссельском столичном регионе. **English MT:** The School of Business and Social Sciences at Aarhus University was founded in 1928 and is a member of the Association of European Universities, headquartered in the Brussels-Capital Region.	0.88	0.56	1.00	0.90	3rd
Graph: 11 Diagonal Street | location | South Africa **Text:** Диагонал-стрит 11 находится в Южной Африке. **English MT:** Diagonal Street 11 is located in South Africa.	0.12	1.00	0.12	1.00	5th

Table 15: Russian samples from 4L-RP-Human with their Human and MonoLR scores in a scale from 0 to 1. The samples were selected from the 1st (good), 3rd (median) and 5th (bad) quintile based on the accuracy of MonoLR compared to the Human judgment.

Sample	Precision		Recall		Quintile
	Human	MonoLR	Human	MonoLR	
Graph: McVeagh of the South Seas | imdbId | 0004319 McVeagh of the South Seas | director | Cyril Bruce McVeagh of the South Seas | director | Harry Carey (actor born 1878) McVeagh of the South Seas | starring | Harry Carey (actor born 1878) McVeagh of the South Seas | writer | Harry Carey (actor born 1878) **Text:** Ysgrifennodd Harry Carey (a anwyd yn 1878) McVeagh of the South Seas a cyfarwyddodd Cyril Bruce. Mae gan y ddata IMDb 0004319. **English MT:** Harry Carey (born 1878) wrote McVeagh of the South Seas and Cyril Bruce directed. The IMDb data has 0004319.	0.88	0.96	0.69	0.65	1st
Graph: University of Burgundy | campus | Dijon Dijon | country | France **Text:** Mae Prifysgol Burgundaidd yn cael ei leoli yn Dijon, Ffrainc. **English MT:** Burgundian University is located in Dijon, France.	1.00	0.76	0.94	0.85	3rd
Graph: Bionico | dishVariation | Honey Bionico | country | Mexico **Text:** Mae Bionico yn amrywiad dysgl o Fecsico sy'n cynnwys mêl. **English MT:** Bionico is a variation of a Mexican dish that includes honey.	1.00	0.36	1.00	0.34	5th

Table 16: Welsh samples from 4L-RP-Human with their Human and MonoLR scores in a scale from 0 to 1. The samples were selected from the 1st (good), 3rd (median) and 5th (bad) quintile based on the accuracy of MonoLR compared to the Human judgment.

G Human Annotation Instructions



Language Annotations
perky-loud-mouse-deer.anvil.app
Guest
Built with anvil
Build web apps for free with Anvil

Semantic Evaluation of Graph-Text Pairs

Textual information can be represented as **Subject-Property-Object** Triples.

For example, the following triple:

Subject	Proepry	Object
Texas	location	United States

Represents this text:

Texas is in the U.S.

We can represent more complex textual information by combining multiple triples.

For example, the following table:

Subject	Proepry	Object
Andrews County Airport	location	Texas
Texas	location	United States

Represents this text:

Andrews County Airport is in Texas, U.S.

The aim of this annotation task is to evaluate the similarity of information in **Table-Text** pairs.

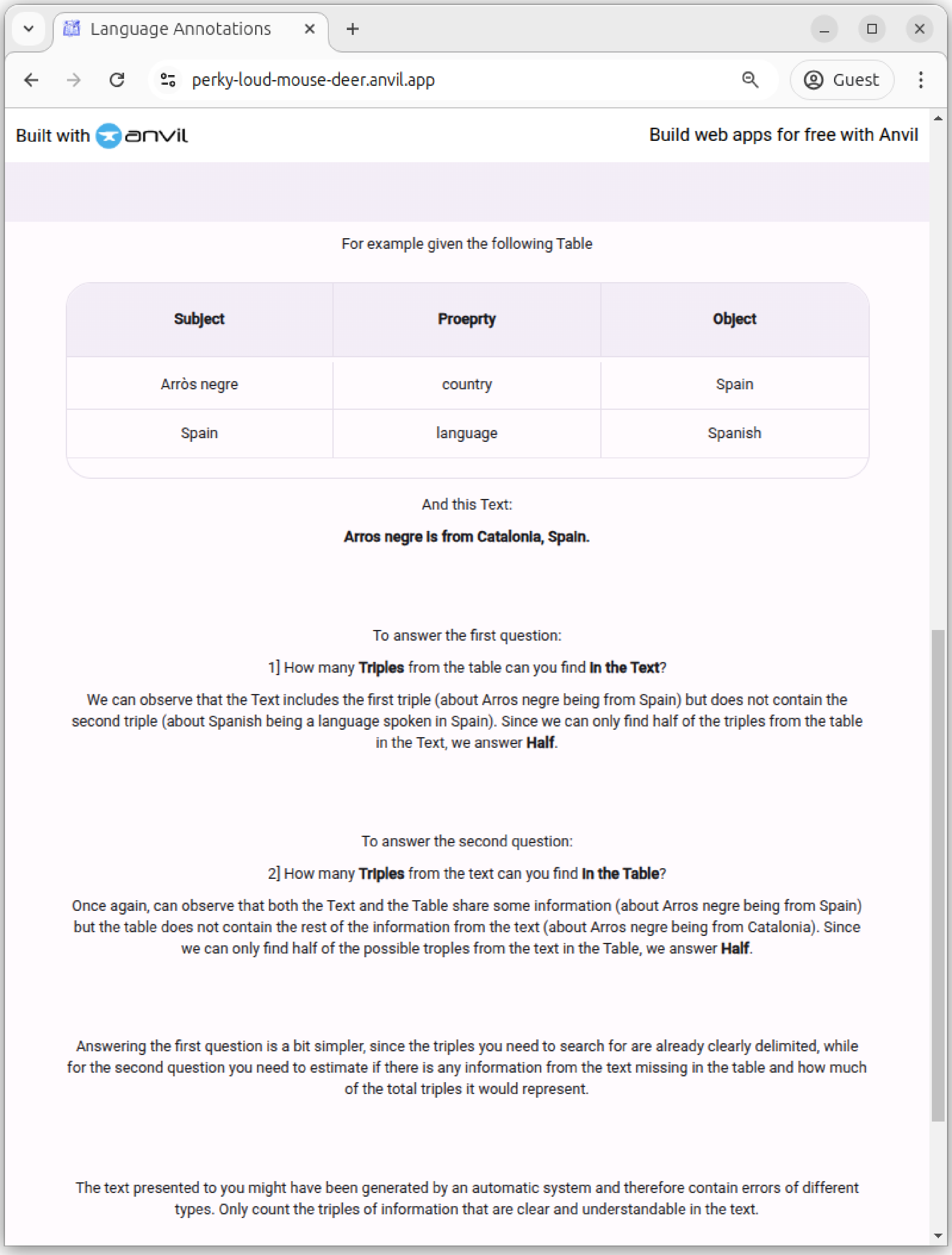
To do so, you will be presented with a **Table** and a **Text** and asked two different questions:

1] How many **Triples** from the table can you find **in the Text**?

2] How many **Triples** from the text can you find **in the Table**?

To answer you will select one of these options: None, Few, Half, Most, All.

Figure 5: Part 1 of 2 from the human annotation instructions.


Language Annotations
perky-loud-mouse-deer.anvil.app
Guest
Built with anvil
Build web apps for free with Anvil

For example given the following Table

Subject	Proeprty	Object
Arròs negre	country	Spain
Spain	language	Spanish

And this Text:

Arros negre is from Catalonia, Spain.

To answer the first question:

1] How many **Triples** from the table can you find **in the Text**?

We can observe that the Text includes the first triple (about Arros negre being from Spain) but does not contain the second triple (about Spanish being a language spoken in Spain). Since we can only find half of the triples from the table in the Text, we answer **Half**.

To answer the second question:

2] How many **Triples** from the text can you find **in the Table**?

Once again, can observe that both the Text and the Table share some information (about Arros negre being from Spain) but the table does not contain the rest of the information from the text (about Arros negre being from Catalonia). Since we can only find half of the possible troples from the text in the Table, we answer **Half**.

Answering the first question is a bit simpler, since the triples you need to search for are already clearly delimited, while for the second question you need to estimate if there is any information from the text missing in the table and how much of the total triples it would represent.

The text presented to you might have been generated by an automatic system and therefore contain errors of different types. Only count the triples of information that are clear and understandable in the text.

Figure 6: Part 2 of 2 from the human annotation instructions.