Phylogeny-Inspired Soft Prompts For Data-to-Text Generation in Low-Resource Languages

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

Yannick Parmentier[1,2]
(1) Université de Lorraine / LORIA
(2) Université d'Orléans / LIFO
yannick.parmentier@loria.fr

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

Abstract

Most work on verbalising Knowledge-Graphs (KG) has focused on high-resource languages such as English, Russian, Czech or Arabic. In this paper, we focus on KG-to-Text generation where the output text is in Breton, Irish or Welsh. To overcome the small size of the parallel training data, we combine the strengths of a multilingual encoder-decoder model with denoising fine-tuning on monolingual data and Soft Prompt fine-tuning on a small quantity of KG/text data. We furthermore structure the soft prompt into multiple sub-prompts designed to capture the similarities and differences between English, Knowledge graphs and the three target languages. Our experiments show that our approach outperforms strong baselines and that all sub-prompts contribute to performance[1].

1 Introduction

Data-to-Text generation includes generating complete and precise natural language descriptions of the information contained in structured data like tables or knowledge graphs (KG). The ever-growing volumes of data generated over time have opened the doors to a large variety of data analysis techniques that can be applied to structured data; however, presenting the outcome of these analyses in a straightforward and easy-to-interpret manner can be complex. Data-to-Text generation facilitates the communication of these outcomes by turning cumbersome data structures into accessible text.

Steady progress has been made on the task of generating text from KG graphs into the English language (Castro Ferreira et al., 2020; Pasricha et al., 2020; Guo et al., 2020b; Kertkeidkachorn and Takamura, 2020) and some advances have taken place in other high-resource languages like Russian (Agarwal et al., 2020; Kasner and Dušek, 2020; Yang et al., 2020). Little research has been done on low-resource languages however which can be explained, at least partially, by how data-intensive the best-performing KG-to-Text approaches are.

Recent work in machine translation (Conneau et al., 2020; Lin et al., 2020) shows that fine-tuning large language models pre-trained on multiple languages helps compensate for data sparsity. Moreover, lightweight fine-tuning techniques have recently emerged that allow preserving the language knowledge obtained from high-resource languages while transferring well to low-resource languages. In particular, the use of phylogeny information has shown good results in transfer learning for related languages in classification tasks like POS tagging, Named Entity Recognition, or Natural Language Inference (Faisal and Anastasopoulos, 2022). At the same time, Factorized Soft Prompts have demonstrated a good performance in transfer learning to low-resource languages in text generation tasks like summarization (Vu et al., 2022).

In this work, we focus on Data-to-Text generation where the input is a Knowledge Graph in the RDF (Resource Description Format, (W3C, 1999)) format and the output is a text verbalising this graph in languages from the Celtic family namely, Irish (GA), Welsh (CY), and Breton (BR). We propose an approach which combines the strengths of large multilingual language models (mT5) with monolingual denoising pre-training and linguistically motivated, lightweight fine-tuning on small quantities (around 1.5K) of downstream RDF-to-Text data. Fine-tuning a multilingual model using mono-lingual denoising incorporates the benefits of the large quantities of unlabeled data, which is particularly important for low-resource languages. We further hypothesize that structuring the Soft Prompt to account for relations between languages helps improve transfer learning.

Leveraging the data made available by the

[1] Code at https://gitlab.inria.fr/wsotomar/phylogenyinspired_softprompts

Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 186–198
November 1–4, 2023. ©2023 Association for Computational Linguistics

WebNLG shared task[2], we show that our approach outperforms Full Model Fine-tuning and Factorized Soft Prompts without phylogeny information in terms of both automatic metrics and human judgments. We also perform an ablation study to study the impact of the various sub-prompts, and we examine how the size (from 0 to 1.5K) of the RDF-to-Text fine-tuning data impacts performance.

2 Related Work

RDF-to-Text. This is a subclass of the Data-to-Text task that takes as input RDF graphs and aims to generate natural text. A usual benchmark for this task is the WebNLG Challenge (Gardent et al., 2017; Castro Ferreira et al., 2020) which uses DBpedia graphs of different sizes (from 1 to 7 triplets) as the sources and includes human written lexicalizations as the targets.

The best-performing approaches for this task are based on the Transformer Architecture (Vasava et al., 2022). Ribeiro et al. (2021) tested the efficiency of LLMs on Graph-to-Text tasks and Li et al. (2020) did it specifically on the WebNLG Challenge. They both found that T5 (Raffel et al., 2020) performs particularly well. Later on, using lightweight approaches like Prefix Tuning (Li and Liang, 2021) and Control Prefixes (Clive et al., 2021) on T5, further improvements were reached. Currently, the best results for the English WebNLG are around **57 BLEU** for all the categories.

Beyond English. Some research has expanded the results obtained in the Data-to-Text task from English to other languages. Agarwal et al. (2020) leverage the strength of pre-trained language models. They further pre-train T5 on parallel English-Russian machine translation data for around 900K steps before fine-tuning on 34K English WebNLG and 29K Russian WebNLG samples. This method obtained a balanced score of around **52 BLEU** for both English and Russian. Their results show the benefits of pre-trained language models, even when the target language is new to the model.

Kale and Roy (2020) attempt the Table-to-Text task in Czech. They pre-train a transformer from scratch on English-to-Czech parallel data for a million steps and then fine-tune it on 1K Table-to-Czech samples. Their best-performing model obtains around **26 BLEU**. They prove that good pre-training can produce acceptable results even when samples of the downstream task are limited.

Demir (2022) experiments with Recurrent Neural Networks, training a Seq2Seq model from scratch for Turkish Data-to-Text. Their best-performing model obtains **31 BLEU** after being trained on close to 40K samples mined from the Turkish Wikipedia.

Touma et al. (2023) fine-tune various models on 7K WebNLG samples translated to Arabic. Their best-performing model reaches **25 BLEU** and consists of an Encoder-Decoder where both components are initialized on AraBERT (Antoun et al., 2020), a Large Language Model pre-trained on 1.3B Arabic words.

These approaches, however, rely on large (hundred of thousands) bilingual or medium-size (several thousand) data-to-text datasets. In contrast, we rely only on monolingual data, which is more easily available for low-resource languages. We then combine denoising pre-training with a linguistically motivated lightweight fine-tuning strategy to overcome the small size (around 1.5K) of the data-to-text train set.

Parameter-Efficient Training and Low-Resource Languages. To combat catastrophic forgetting and minimise the computation costs induced by the full fine-tuning of large pre-trained models, various parameter-efficient training approaches have surfaced which rely on keeping the original pre-trained model frozen and only training a few additional parameters. In particular, Adapters (parameters introduced in every transformer layer) and Soft Prompts (parameters prepended to the embedded input of the model) have shown good performance on a variety of NLP tasks (Houlsby et al., 2019; Lester et al., 2021).

Parameter-efficient training strategies have also been shown to support transfer learning, which is particularly important when dealing with low-resource languages. Artetxe et al. (2020) showed that it is possible to adjust an existing language model to a new language using Adapters. Pfeiffer et al. (2020) demonstrated that stacking tasks and target language adapters can be used for multilingual transfer learning of a task from a high-resource language to a low-resource one. In the same line, Vu et al. (2022) showed that by using Factorized Soft Prompts that separate tasks from the target language it is possible to transfer learning of generative tasks like text summarization from

[2] https://synalp.gitlabpages.inria.fr/webnlg-challenge/challenge_2023/

high-resource languages to low-resource languages like Vietnamese and Thai. Lee et al. (2022) showed that fusing Language Family adapters improves performance in low-resourced languages. Faisal and Anastasopoulos (2022) showed that training hierarchical language adapters following a phylogeny tree during training can further improve the transfer learning capacity of adapters in classification tasks like POS Tagging, Name Entity Recognition and Natural Language Inference.

We build on these approaches and extend Vu et al. (2022)'s Soft Prompt approach by structuring the prompt to better account for the phylogeny relations between languages and maximise transfer learning between closely related languages.

3 Phylogeny-Inspired Task-Source-Target Soft Prompts

At the heart of our approach is a highly structured Soft Prompt which decomposes into multiple sub-prompts. Inspired partly by the structure of the original T5 translation prompts (Raffel et al., 2020) (e.g., *Translate English to German*), we first divide the Soft Prompt into three main components: Task, Source, and Target. This is also similar to one standard practice in Machine Translation architectures like mBART (Liu et al., 2020), M2M100 (Fan et al., 2021), and NLLB (NLLB Team et al., 2022) where both Source and Target languages are specified to improve Zero-Shot performance.

In an attempt to model phylogeny information, we further decompose the Source and Target components into Family, Genus, and Language sub-prompts. We call the resulting Soft Prompt, "Phylogeny-Inspired Task-Source-Target" Soft Prompt (PI-TST). By using this prompt, we aim to allow less-resourced languages to benefit from the training data of their related languages while preventing the mixture of training data to introduce too much noise to the model. Figure 1 shows the simplified phylogeny tree we used during training. We paired the linearized RDFs with English since the subjects, objects, and predicates of the RDF are in English.

4 Method

We fine-tune a pre-trained multilingual model using the Soft Prompt described in the previous section and proceed in two steps: an unsupervised pre-training of the whole Soft Prompt (Step 1) and a downstream task fine-tuning of the task sub-prompt (Step 2). During these two steps, all the weights of the base model remain frozen. For mT5, we add a preliminary step (Step 0) which we refer to as language Model Adaptation.

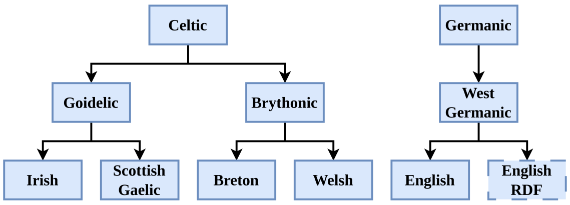


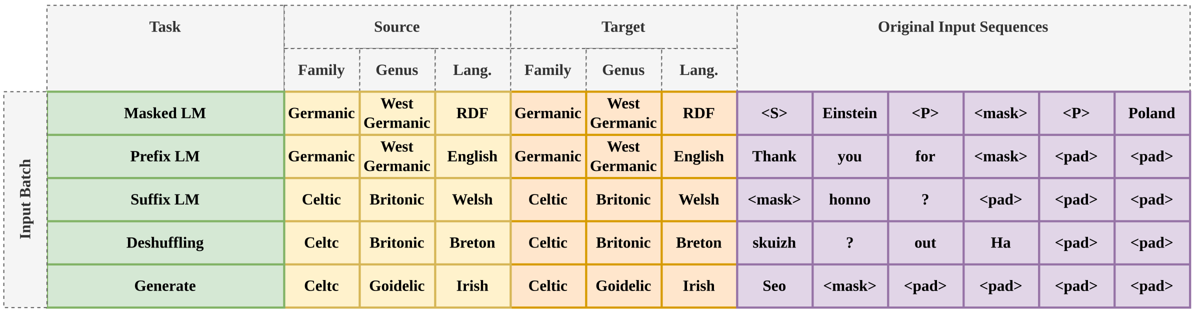
Figure 1: Simplified phylogeny tree used when training the Soft Prompts.

Step 0: Language Model Adaptation. Sometimes the pre-training objective of the base Large Language Model is not aligned with the natural text generation objective. For example, models based on T5 are generally pre-trained on the Span Corruption objective which generates spans of text separated by sentinel tokens instead of plain natural text. When performing full model fine-tuning this behaviour is soon corrected, but for lightweight approaches like Soft Prompts, it can be harder to overcome it. Lester et al. (2021) proved that pre-training the base model for some steps in a language modelling task, like Prefix Language Modelling, before freezing it and applying a lightweight strategy benefits performance. We use the BERT-style Masked Language Modelling (MLM) pre-training task of Raffel et al. (2020) instead of Prefix Language Modelling (PLM). We do this given the better performance of the first objective over the second in Raffel et al. (2020), particularly on translation downstream tasks. Furthermore, the MLM task is closer to our downstream tasks than PLM. Once this step has been completed, we freeze the base model for the rest of the training.

Step 1: Unsupervised Pre-training of the Soft Prompt. The goal of the first stage is to train the language components of the Soft Prompt so that each of them captures as much language information relevant to their assigned language. Specifically, we train the whole Soft Prompt on a mixture of unsupervised, monolingual tasks (Masked LM, Prefix LM, Suffix LM, Generation, and Deshuffling). We substitute the parameters being used for each component based on the language of the training sample. Instances that belong to the same language family share the same Family sub-prompt but have different Genus and Language

Soft Prompt Component	Possible Options
Task	Masked LM, Prefix LM, Suffix LM, Deshuffling, Open Generation, Data-to-Text
Source/Target Family	Germanic, Celtic
Source/Target Genus	West Germanic, Goidelic, Britonic
Source/Target Language	English, RDF, Irish, Scottish Gaelic, Breton, Welsh

Table 1: **Possible values of each Soft Prompt component.**



	Task	Source			Target			Original Input Sequences					
		Family	Genus	Lang.	Family	Genus	Lang.						
Input Batch	Masked LM	Germanic	West Germanic	RDF	Germanic	West Germanic	RDF	<S>	Einstein	<P>	<mask>	<P>	Poland
	Prefix LM	Germanic	West Germanic	English	Germanic	West Germanic	English	Thank	you	for	<mask>	<pad>	<pad>
	Suffix LM	Celtic	Britonic	Welsh	Celtic	Britonic	Welsh	<mask>	honno	?	<pad>	<pad>	<pad>
	Deshuffling	Celtc	Britonic	Breton	Celtic	Britonic	Breton	skuizh	?	out	Ha	<pad>	<pad>
	Generate	Celtc	Goidelic	Irish	Celtic	Goidelic	Irish	Seo	<mask>	<pad>	<pad>	<pad>	<pad>

Figure 2: Example input Batch for Step 1 (Unsupervised Pre-training of the Soft Prompt).

sub-prompts. Table 1 shows the possible values of each component and Figure 2 shows an example input batch for this step.

Step 2: Downstream Task Fine-tuning of the Soft Prompt. Once the language components of the Soft Prompt have learned to perform the unsupervised tasks, we freeze them and train the Task sub-prompt on the downstream task (RDF-to-Text generation). Following Vu et al. (2022), we use one of the unsupervised task Soft Prompt components to initialize the new task Soft Prompt component. In our case, we use the Masked LM component since we consider it to be the closest one to the RDF-to-Text task. In this stage, we continue switching the language components of the Soft Prompt as required by each training instance, with the difference that now they are frozen.

Inference. At inference time all the parameters of the base model and the Soft Prompt remain frozen. We then combine the task and language sub-prompts as required by each of the 3 inference tasks (i.e., generating into Breton, Irish or Welsh).

5 Data

Table 2 shows the number of samples available on each dataset used.

For unsupervised training, we extract Celtic and English monolingual data from multiple datasets available in the Huggingface Hub [3]. Specifically, we collected data from different OPUS corpora (Tiedemann, 2012) (Bible Corpus (Christodouloupoulos and Steedman, 2015), DGT, EUConst, GNOME, KDE4, OfisPublik (Tyers, 2009), OpenSubtitles (Lison and Tiedemann, 2016), Opus-100 (Zhang et al., 2020), ParaCrawl, QED (Abdelali et al., 2014), Tatoeba, and Ubuntu), CC-100 (Conneau et al., 2020), CC-Aligned (El-Kishky et al., 2020), CC-Matrix (Schwenk et al., 2021), ECDC Steinberger et al. (2014), mC4, OSCAR (Suárez et al., 2019), TaPACo (Scherrer, 2020), TedTalks (Cettolo et al., 2012), UDHR, and Wikipedia.

Version	Train Sample	Validation Samples	Test Samples
Monolingual			
BR	1 206 546	250	250
CY	12 993 205	250	250
EN	7 959 035	250	250
GA	7 996 721	250	250
GD	1 019 593	250	250
WebNLG			
RDF-to-BR	–	1 399	2 280
RDF-to-CY	–	1 665	1 779
RDF-to-EN	35 426*	4 464	5 150
RDF-to-GA	–	1 665	1 779

Table 2: **Datasets.** Collected data for the experiment. While a big dataset of monolingual data was collected, just a small amount is seen during training. *The training WebNLG data was only used during the Zero-Shot ablation experiment.

To process the text we first split it into sentences

[3] https://huggingface.co/datasets

using SentenceSplitter [4] with the default English settings. Then, each sentence was normalized using TextaCy[5], we applied bullet point normalization, hyphenated words normalization, quotation marks normalization, Unicode normalization, white space normalization, and HTML tag removal. Finally, the sentences were filtered using FastText Language Identification (Joulin et al., 2016b,a)[6] by keeping only those above a 0.5 threshold. We collected as many samples as possible for the Celtic languages but limited the number of English samples to prevent it from overshadowing the other languages.

6 Experimental Setup

6.1 Training Process

Step 0: Language Model Adaptation. We choose to use mT5-Large (Xue et al., 2021) as our base model[7] since it that was originally pretrained in several languages including English, Irish, Scottish Gaelic, and Welsh. Before training the Phylogeny-Inspired Soft Prompt we perform a language model adaptation for 30 000 steps on monolingual data for English, Breton, Irish, Scottish Gaelic, and Welsh as well as RDF triples from WebNLG. Once the LM Adaptation has been completed the base model is permanently frozen and we train the Phylogeny-Inspired Soft Prompts.

Step 1: Unsupervised Pre-training of the Soft Prompt. We perform this step for 30 000 steps over our monolingual data for English, Breton, Irish, Scottish Gaelic, and Welsh as well as the RDF triples from WebNLG.

Step 2: Downstream Task Fine-tuning of the Soft Prompt. We fine-tune the Task sub-prompt on the WebNLG task using the validation split of the English WebNLG dataset (Gardent et al., 2017) as well as human-written Breton, Irish, and Welsh translations of it. This process takes 5 epochs or around 4500 steps and we keep the best checkpoint every 500 steps.

To account for the unbalanced distribution of samples in our datasets we apply the sampling strategy described in Devlin et al. (2019) with $\alpha = 0.3$ which has been shown to perform best NLLB Team et al. (2022). Table 3 accounts for that and other relevant hyperparameters used. The batch size was chosen to optimize the use of our GPUs. The learning rate was chosen after a small exploratory experiment. The Soft Prompt size follows Vu et al. (2022) using around 50 tokens for task and 50 for each language. Finally, the training steps follow Lester et al. (2021).

Parameter	Value
Base Model	mT5-Large
Vocabulary Size	∼250K Tokens
Embedding Dimensions	1 024
Base Model Parameters	∼1.22B
Total Prompt Parameters	∼747K
Inference Prompt Parameters	∼143
Learning Rate	0.0001
Batch Size per GPU	8
Available GPUS	2 Nvidia A40
Sampling Temperature	0.3
ML Adaptation Steps	30 000
ML Adaptation Training Hours	∼12
Soft Prompt Pre-training Steps	30 000
Soft Prompt Pre-training Training Hours	∼12
Soft Prompt Fine-tuning Steps	∼4 500
Soft Prompt Fine-tuning Training Hours	∼4

Table 3: **Hyperparameters.**

6.2 Models

We compare our model to a baseline obtained by applying full fine-tuning on mT5, to previous work, and two MT-based, upper-bound models.

Full Model Fine-tuning. We perform full fine-tuning on mT5. First, we performed the Language Model Adaptation to attune the model to our target languages. We then fine-tuned it on the downstream task.

Control Prefixes. The Control Prefixes model presented by Clive et al. (2021) is currently one of the best-performing strategies for the English WebNLG benchmarks. This lightweight fine-tuning approach includes attribute-level parameters into different layers of T5 which indicate the semantic category of the input WebNLG RDF graph to improve performance. For our baseline, we trained Control Prefixes on the WebNLG validation data of all languages (Celtic and English).

Machine Translation (MT). We consider two scenarios using Machine Translation: a generate-and-translate scenario (NLG+MT), where the output of the best RDF-to-English generation system from the WebNLG Challenge 2020 (Guo et al., 2020a) is translated into the Celtic languages using

[4] https://github.com/mediacloud/sentence-splitter

[5] https://textacy.readthedocs.io/en/latest/

[6] https://fasttext.cc/docs/en/language-identification.html

[7] https://huggingface.co/google/mt5-large

Machine Translation, and a translation-only scenario (Gold+MT) where the translation takes as input the references of the WebNLG dataset. We view these models as upper bounds since, different from our models which are trained on around 1.5K data points, the machine translation models have been trained on thousands of samples of parallel English-Celtic data. Note further that the GOLD+MT model does not perform RDF-to-Text generation as it simply translates the English sentences of the WebNLG test set into Celtic. To perform the translations we used a version of the system from Zhang et al. (2020) trained only on Celtic and English data from the OPUS Corpora (Tiedemann, 2012). It is worth noting that NLG+MT and the Gold+MT models are requires significantly more parallel data to be trained than our proposed method.

6.3 Ablation Experiments

To test the impact of the various sub-prompts (task, phylogeny data, source and target language), we perform a series of ablation experiments. Figure 3 shows the various prompts we experiment with. We compare our full prompt with five other prompts: the same prompt but without phylogeny information (TST); the same prompt without Source Language information (PI-TT) and three simpler prompts without phylogeny information which either are unstructured (S) or model only two factors namely, Task and Target Language (TT) or Source and Target Language (ST).

We fixed the size of the Soft Prompts at 140 tokens for all the experiments. When a task component was present, we fixed its size to 50 tokens with the rest taking 90 tokens. All the language-related components on a Soft Prompt had their size distributed uniformly as shown in Figure 3. All the Soft Prompts underwent the same pre-training before the downstream task fine-tuning.

6.4 Training Data

Training Samples. This experiment tests our final PI-TST model but fine-tuned on different numbers (100, 500, 1000) of randomly sampled elements from each language on the dataset.

Zero-Shot. We test the zero-shot capabilities of our final PI-TST model by fine-tuning the task Soft Prompt only in English (either on the validation or training data) and testing it on Celtic languages.

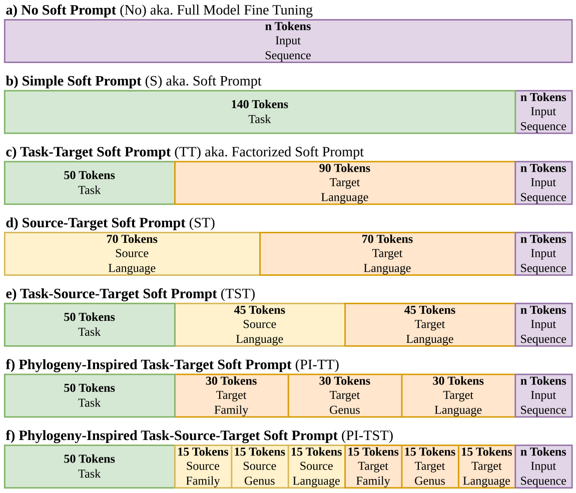

a) No Soft Prompt (No) aka. Full Model Fine Tuning
n Tokens Input Sequence
b) Simple Soft Prompt (S) aka. Soft Prompt
140 Tokens Task
n Tokens Input Sequence
c) Task-Target Soft Prompt (TT) aka. Factorized Soft Prompt
50 Tokens Task
90 Tokens Target Language
n Tokens Input Sequence
d) Source-Target Soft Prompt (ST)
70 Tokens Source Language
70 Tokens Target Language
n Tokens Input Sequence
e) Task-Source-Target Soft Prompt (TST)
50 Tokens Task
45 Tokens Source Language
45 Tokens Target Language
n Tokens Input Sequence
f) Phylogeny-Inspired Task-Target Soft Prompt (PI-TT)
50 Tokens Task
30 Tokens Target Family
30 Tokens Target Genus
30 Tokens Target Language
n Tokens Input Sequence
f) Phylogeny-Inspired Task-Source-Target Soft Prompt (PI-TST)
50 Tokens Task
15 Tokens Source Family
15 Tokens Source Genus
15 Tokens Source Language
15 Tokens Target Family
15 Tokens Target Genus
15 Tokens Target Language
n Tokens Input Sequence

Figure 3: Composition of the input embeddings to the frozen model for different Soft Prompt variants.

7 Evaluation

7.1 Automatic Evaluation

BLEU. We compute the corpus level BLEU score (Papineni et al., 2002) for each experiment using SacreBLEU (Post, 2018) .

Google BLEU. We also compute sentence level Google BLEU scores (Wu et al., 2016) for each experiment.

LaBSE Cosine Similarity. We use LaBSE (Feng et al., 2022) to obtain sentence embeddings for the generated text and the human reference. We then compute the sentence level cosine similarity of both embeddings as an automatic measurement of semantic accuracy. We choose this model for sentence embeddings over others given its implicit goal of being language agnostic, which benefits the experimentation with low-resourced languages.

Wilcoxon signed-rank test. We use the Wilcoxon signed-rank test (Wilcoxon, 1945) on the sentence level metrics (Google BLEU and LaBSE Cosine Similarity) to evaluate if the differences observed on different experiments are statistically significant. We proffered this approach over paired Student's t-test since our results do not follow a normal distribution.

7.2 Human Evaluation

We selected 25 random input graphs from our test set making sure to have a variety of sizes and collected the generation by our proposed model from those graphs into all our target languages. We then provided all 25 of those generated texts to human evaluators and asked them to score them following 4 different criteria: Readability, Grammaticality, Word Order and Semantic Adequacy. Each criterion was to be scored on a 1 to 3 scale where 1 is bad, 2 is medium, and 3 is good.

Readability. The evaluator was given the generated output of the model and asked if the generated text was understandable and reasonable text in the language.

Grammaticality. The evaluator was given the generated output of the model and asked if the morphology of the generated text was correct and if agreement constraints (e.g., verb/subject, noun/adjective) were respected.

Word Order. The evaluator was given the generated output of the model and asked if the word order of the generated text was correct and if a native speaker would come up with a text like that.

Semantic Adequacy. The evaluator was given the generated output of the model as well as the human-written reference and asked if the generated text shared the same meaning as the human-written reference.

We reached out to colleagues that grew up on regions where the evaluated language is spoken to perform the human evaluation. Given the nature of the low-resource languages we are working with, we only collected a small number of evaluations.

8 Results

8.1 Automatic Evaluation Results

PI-TST outperforms the baselines. Table 4 shows the results of the automatic evaluation. Our proposal (PI-TST) outperforms mT5 Full Fine-tuning and the state-of-the-art, Control Prefixes models fine-tuned on Celtic. For Breton and Welsh, PI-TST even outperforms the BLEU score of the NLG+MT approach, with the advantage that our model does not require any number of parallel translation data, while the MT model has been trained on significant amounts of bilingual data, which is not always available for low resource languages. Furthermore, the NLG model of the NLG+MT baseline was trained on all 32K samples of the full English WebNLG while PI-TST is only trained on validation data, which is significantly smaller. It is worth noting that, for Breton, which is the most under-resourced of the Celtic language evaluated, our method even comes close to the Gold+MT BLEU score and surpasses its LaBSE Cosine Similarity score. As the data used to pretrain mT5 does not include any Breton, this suggests that our fine-tuning approach produces bigger improvements on languages which were not seen during the base model pre-training.

The effect of Source information. The ablation results in Table 5 show that the two best-performing models (PI-TST, ST) include a source sub-prompt, which suggests that, similar to the control tokens used in multilingual machine translation, our source and target sub-prompts help structure the representation space and guide learning. We conjecture that having both Source and Target sub-

	BLEU Score (↑)				Google BLEU Score (↑)*				LaBSE Cosine Similarity (↑)			
Experiment	BR	CY	EN	GA	BR	CY	EN	GA	BR	CY	EN	GA
Machine Translation												
NLG+MT	*13.08*	*20.24*	*53.98*	*18.09*	*17.74*	*27.49*	*49.64*	*24.86*	*72.96*	*89.90*	*95.05*	*87.76*
Gold+MT	*19.81*	*49.04*	*100.00*	*32.09*	*23.04*	*51.82*	*100.00*	*36.44*	*76.23*	*94.80*	*100.00*	*92.56*
Baselines												
Control Prefixes	12.23	13.33	**51.61**	8.17	16.37	18.76	47.77	13.59	80.52	79.41	**94.52**	73.12
Full Fine-tuning	16.49	18.83	46.40	14.16	21.36	24.36	43.62	20.09	82.56	86.02	92.35	82.49
Final												
PI-TST	**18.15**	**20.60**	49.15	**15.64**	**22.57**	**25.95**	46.09	**21.23**	**84.09**	**87.72**	93.65	**84.68**

Table 4: **Automatic Evaluation Results.** For Google BLEU and Cosine Similarity, the results without a statistically significant difference from the final PI-TST model (p > 0.05) are underlined. The English values on the Machine Translation rows are the scores obtained by the RDF-to-EN model and the Gold references i.e., in this case, translation is not used. *Since we use the Sentence Level Google BLEU score for statistical significance analysis, here we present the Average of the Sentence level scores instead of the corpus level one.

	BLEU Score (↑)				Google BLEU Score (↑)*				LaBSE Cosine Similarity (↑)			
Experiment	BR	CY	EN	GA	BR	CY	EN	GA	BR	CY	EN	GA
Soft Prompt												
S	9.63	1 1.01	48.48	10.36	13.41	15.18	44.73	14.18	79.84	86.42	93.51	82.49
TT	17.70	19.94	48.30	15.58	21.95	25.32	45.26	21.04	83.21	87.59	93.60	84.66
ST	17.89	19.94	49.18	15.58	22.24	25.34	45.73	20.88	83.72	87.53	93.55	84.47
TST	16.28	18.49	47.29	15.39	21.33	24.19	44.82	20.94	82.21	86.46	93.04	84.16
PI-TT	17.43	19.41	48.32	15.23	22.16	25.28	45.29	**21.48**	83.55	87.34	92.90	84.35
Training Samples												
100 Samples	12.42	13.61	38.42	10.66	17.12	18.98	38.09	15.66	77.58	81.15	89.74	78.56
500 Samples	14.31	14.95	43.70	12.60	19.08	20.68	42.12	16.99	79.92	82.54	91.30	79.84
1000 Samples	15.34	18.29	47.18	13.91	20.29	24.02	44.29	19.32	81.99	86.44	92.47	82.53
Zero-Shot												
English Validation	9.81	11.85	48.36	9.69	13.79	16.88	45.04	13.96	78.57	83.29	93.26	82.19
English Training	9.57	11.27	48.09	10.36	13.49	16.19	44.95	14.58	79.19	83.70	92.94	81.04
Final												
PI-TST	**18.15**	**20.60**	**49.15**	**15.64**	**22.57**	**25.95**	**46.09**	21.23	**84.09**	**87.72**	**93.65**	**84.68**

Table 5: **Automatic Evaluation Results of Ablation Experiments.** For Google BLEU and Cosine Similarity, the results without a statistically significant difference from the final PI-TST model (p > 0.05) are underlined. *Since we use the sentence level Google BLEU score for statistical significance analysis, here we present the average of the sentence level scores instead of the corpus level one.

prompts (rather than just Target) helps the model differentiate between the unsupervised monolingual step (Step 1) where Source and Target prompts refer to the same language and the second fine-tuning step where the Source and Target prompt refers to different languages (Source: RDF, Target: Celtic). On the other hand, we observe that the TST model has much lower performance than ST, which is likely due to a trade off between prompts and prompt size: 70 tokens for the Source token in ST vs. 45 in TST.

The effect of Phylogeny information. Just like PI-TST, the Phylogeny-Inspired Task-Target (PI-TT) model outperforms Full Model Fine-tuning in all languages confirming the positive impact of phylogeny information.

Languages not seen during pre-training of the original Encoder-Decoder (mT5). For Breton, the only language not seen during the pretraining of mT5, the PI-TT model outperforms TT indicating that phylogeny information is particularly useful for under-resourced languages.

Source and Phylogeny Prompts. Comparing models across these two dimensions, we find that while adding either a phylogeny or a source subprompt does not always improve performance (both TST and PI-TT underperform TT), adding both does (PI-TST outperforms all other models).

Size of the Training Data. Figure 4 shows the performance of the PI-TST models when fine-tuned with varying amounts of KG-Text data. With only 1 000 samples per language, PI-TST outperforms Full Model fine-tuning in English and performs on par with the Celtic languages.

Zero-Shot. Table 5 shows that using our model in a zero-shot setting reaches equivalent results on Celtic languages than a simple Soft Prompt model trained on all Celtic languages.

Statistical Significance. Table 7 in Appendix A presents the statistical significance between each experiment and our final proposal PI-TST. While some of the ablation experiments produce results that are not statistically different to our proposal, we still advocate from our proposal over those other approaches, since PI-TST provides much more controlability and flexibility given its complex soft prompt. We believe that the extreme modularity of our proposal gives it an edge over the ablation studies. We also note that, where the average Google BLEU score of an ablation experiments outperformed our model (Irish PI-TT) the difference was not statistically significant. Finally, the difference on the Google BLEU score between our proposal and the Breton Gold+MT is not statistically significant; despite the former (and more data intensive) approach having a higher average.

8.2 Human Evaluation Results

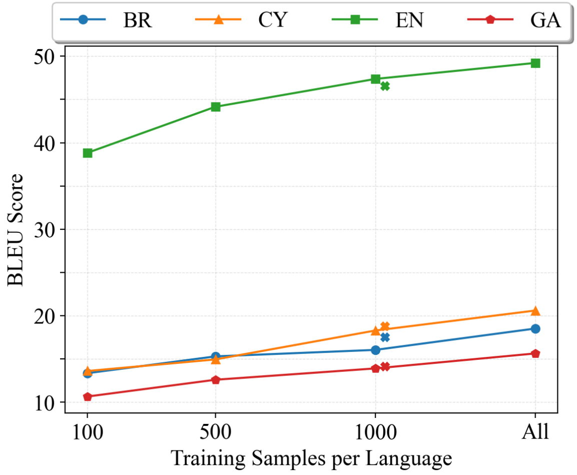


Figure 4: BLEU Score comparison by number of training samples per language. The × mark indicates the score of Full Model Fine-tuning.

When asked where they learned the language 4 of the evaluators answered "Home", 2 answered "School and Home" and 3 answered "School". When asked how they considered their proficiency at the language 8 of the evaluators answered "Good" and 1 answered "Medium". Table 6 shows the results of their evaluation of our PI-TST model.

This evaluation shows that the model produces acceptable text concerning Readability, Grammaticality and Word Order for all Celtic languages. It also shows that, for English and Irish, the quality of the Semantic Adequacy is past the middle point.

Criteria (↑)	BR	CY	EN	GA
Annotators (↑)	3	2	2	2
Readability (↑)	2.67	2.18	2.96	2.16
Grammaticality (↑)	2.69	2.46	2.84	2.42
Word Order (↑)	2.68	2.58	2.94	2.30
Semantic Adequacy (↑)	1.84	1.64	2.54	2.06

Table 6: **Results of Human Evaluation**

9 Conclusion

In this work, we proposed a Soft Prompt approach enriched with phylogeny source language information. We showed that adding this information to the Soft Prompt leads to an improvement in the Data-to-Text task on low-resource languages. In particular, we showed that this approach can outperform basic strategies like Full Model Fine-tuning and other complex approaches like Control Prefixes, simple Soft Prompts, and Factorized Soft Prompts. These results open the door to further advancements in the NLG domain for low-resource languages, as shown by the improved performance of Breton which was new to the base language model and is significantly less resourced than the other Celtic languages studied.

10 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`).

11 Limitations

The scarcity of available data and the access to native speakers of the language made the research particularly challenging. Furthermore, we only tested our approach in one language Family given the lack of training and evaluation data in other language families. We would like to expand our approach to other generative tasks and cover more language families across the globe.

12 Ethics Statement

Research into expanding the capabilities of advanced NLP tools to low-resource languages facilitates the democratization of these technologies. Particularly in the domain of Data-to-Text generation, this research can be used to make data available to speakers of low-resourced languages. However, it is important to consider the implications and shortcomings of these technologies. As happens in high-resourced languages, current models are still capable of generating inaccurate text and misleading users.

References

Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, and Stephan Vogel. 2014. The AMARA corpus: Building parallel language resources for the educational domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1856–1862, Reykjavik, Iceland. European Language Resources Association (ELRA).

Oshin Agarwal, Mihir Kale, Heming Ge, Siamak Shakeri, and Rami Al-Rfou. 2020. Machine translation aided bilingual data-to-text generation and semantic parsing. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 125–130, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

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.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Annual conference of the European Association for Machine Translation*, pages 261–268, Trento, Italy. European Association for Machine Translation.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Language resources and evaluation*, 49:375–395.

Jordan Clive, Kris Cao, and Marek Rei. 2021. Control prefixes for text generation. *CoRR*, abs/2110.08329.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Seniz Demir. 2022. Turkish data-to-text generation using sequence-to-sequence neural networks. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(2):1–27.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969, Online. Association for Computational Linguistics.

Fahim Faisal and Antonios Anastasopoulos. 2022. Phylogeny-inspired adaptation of multilingual models to new languages. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 434–452, Online only. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond English-centric multilingual machine translation. *The Journal of Machine Learning Research*, 22(1):4839–4886.

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.

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, Ning Dai, Xipeng Qiu, Xiangyang Xue, David Wipf, and Zheng Zhang. 2020a. $\mathscr{P}^2$: A plan-and-pretrain approach for knowledge graph-to-text generation. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 100–106, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Qipeng Guo, Zhijing Jin, Xipeng Qiu, Weinan Zhang, David Wipf, and Zheng Zhang. 2020b. 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.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016a. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016b. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Mihir Kale and Scott Roy. 2020. Machine translation pre-training for data-to-text generation - a case study in Czech. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 91–96, Dublin, Ireland. Association for Computational Linguistics.

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.

Natthawut Kertkeidkachorn and Hiroya Takamura. 2020. Text-to-text pre-training model with plan selection for RDF-to-text generation. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 159–166, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Jaeseong Lee, Seung-won Hwang, and Taesup Kim. 2022. FAD-X: Fusing adapters for cross-lingual transfer to low-resource languages. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 57–64, Online only. Association for Computational Linguistics.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Xintong Li, Aleksandre Maskharashvili, Symon Jory Stevens-Guille, and Michael White. 2020. Leveraging large pretrained models for WebNLG 2020. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 117–124, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2020. Pretraining multilingual neural machine translation by leveraging alignment information. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2649–2663, Online. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pretraining for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

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. *arXiv preprint arXiv:2207.04672*.

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.

Nivranshu Pasricha, Mihael Arcan, and Paul Buitelaar. 2020. NUIG-DSI at the WebNLG+ challenge: Leveraging transfer learning for RDF-to-text generation.

In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 137–143, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. 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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Leonardo F. R. Ribeiro, Martin Schmitt, Hinrich Schütze, and Iryna Gurevych. 2021. Investigating pretrained language models for graph-to-text generation. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 211–227, Online. Association for Computational Linguistics.

Yves Scherrer. 2020. TaPaCo: A corpus of sentential paraphrases for 73 languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6868–6873, Marseille, France. European Language Resources Association.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.

Ralf Steinberger, Mohamed Ebrahim, Alexandros Poulis, Manuel Carrasco-Benitez, Patrick Schlüter, Marek Przybyszewski, and Signe Gilbro. 2014. An overview of the European Union's highly multilingual parallel corpora. *Language Resources and Evaluation*, 48(4):679–707.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Roudy Touma, Hazem Hajj, Wassim El-Hajj, and Khaled Shaban. 2023. Automated generation of human-readable natural Arabic text from rdf data. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(4):1–13.

Francis M. Tyers. 2009. Rule-based augmentation of training data in Breton-French statistical machine translation. In *Proceedings of the 13th Annual conference of the European Association for Machine Translation*, Barcelona, Spain. European Association for Machine Translation.

Himil Vasava, Pramegh Uikey, Gaurav Wasnik, and Raksha Sharma. 2022. Transformer-based architecture for empathy prediction and emotion classification. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 261–264, Dublin, Ireland. Association for Computational Linguistics.

Tu Vu, Aditya Barua, Brian Lester, Daniel Cer, Mohit Iyyer, and Noah Constant. 2022. Overcoming catastrophic forgetting in zero-shot cross-lingual generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9279–9300, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

W3C. 1999. Resource description framework (rdf) model and syntax specification. Technical report, W3C. Https://www.w3.org/TR/1999/REC-rdf-syntax-19990222/.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. biom bull 1 (6): 80–83.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Zixiaofan Yang, Arash Einolghozati, Hakan Inan, Keith Diedrick, Angela Fan, Pinar Donmez, and Sonal Gupta. 2020. Improving text-to-text pre-trained models for the graph-to-text task. In *Proceedings of the 3rd International Workshop on Natural Language*

Generation from the Semantic Web (WebNLG+), pages 107–116, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. 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.

A Statistical Significance

Table 7 presents the statistical significance between each experiment and our final proposal PI-TST. While some of the ablation experiments produce results that are not statistically different to our proposal, we still advocate from our proposal over those other approaches, since PI-TST provides much more controlability and flexibility given its complex soft prompt. We believe that the extreme modularity of our proposal gives it an edge over the ablation studies. We also note that, where the average Google BLEU score of an ablation experiments outperformed our model (Irish PI-TT) the difference was not statistically significant. Finally, the difference on the Google BLEU score between our proposal and the Breton Gold+MT is not statistically significant; despite the former (and more data intensive) approach having a higher average.

	Google BLEU Score (↑)*				LaBSE Cosine Similarity (↑)			
Experiment	BR	CY	EN	GA	BR	CY	EN	GA
Machine Translation								
NLG+MT	0.0000	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Gold+MT	0.2700	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Baselines								
Control Prefixes	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Full Fine-tuning	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Soft Prompt								
S	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0121	0.0000
TT	0.0007	0.0089	0.0135	0.1101	0.0060	0.4284	0.5420	0.4962
ST	0.0249	0.0048	0.1443	0.0124	0.0626	0.1318	0.4616	0.0467
TST	0.0000	0.0000	0.0000	0.0089	0.0000	0.0000	0.0000	0.0000
PI-TT	0.0166	0.0011	0.0020	0.1358	0.0013	0.0003	0.0000	0.0313
Training Samples								
100 Samples	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
500 Samples	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1000 Samples	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Zero-Shot								
English Validation	0.0000	0.0000	0.0004	0.0000	0.0000	0.0000	0.0000	0.0000
English Training	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000

Table 7: **Wilcoxon signed-rank test p-values.** For Google BLEU and Cosine Similarity, the results without a statistically significant difference from the final PI-TST model (p > 0.05) are underlined. *Since we use the sentence level Google BLEU score for statistical significance analysis, here we present the average of the sentence level scores instead of the corpus level one.