Little Red Riding Hood Goes Around the Globe: Crosslingual Story Planning and Generation with Large Language Models

Evgeniia Razumovskaia[1*]**, Joshua Maynez**[2]**, Annie Louis**[2]**,**
Mirella Lapata[2]**, Shashi Narayan**[2]

[1] Language Technology Lab, University of Cambridge
[2] Google DeepMind

er563@cam.ac.uk, joshuahm@google.com, annielouis@google.com,
lapata@google.com, shashinarayan@google.com

Abstract

Previous work has demonstrated the effectiveness of planning for story generation exclusively in a monolingual setting focusing primarily on English. We consider whether planning brings advantages to automatic story generation across languages. We propose a new task of crosslingual story generation with planning and present a new dataset for this task. We conduct a comprehensive study of different plans and generate stories in several languages, by leveraging the creative and reasoning capabilities of large pretrained language models. Our results demonstrate that plans which structure stories into three acts lead to more coherent and interesting narratives, while allowing to explicitly control their content and structure.

Keywords: story generation, large language models, crosslingual generation, dataset, planning

1. Introduction

Automated story generation has met with fascination since the early days of artificial intelligence. Initial story generation systems required substantial knowledge engineering to create symbolic domain models that described legal characters and their actions, and have almost ubiquitously relied on symbolic planning (Meehan, 1977; Lebowitz, 1987; Riedl and Young, 2010; Ware and robert Michael Young, 2010; Cavazza et al., 2003; Liu and Singh, 2002) and case-based reasoning (Pérez y Pérez and Sharples, 2001; Peinado and Gervás, 2005; Turner, 1992).

The advent of large pre-trained language models (PLMs; Raffel et al. 2020; Lewis et al. 2020; Brown et al. 2020; Chowdhery et al. 2022) has provided a common framework for AI story generation, eschewing the need for manual knowledge engineering. Despite their ability to produce relatively fluent and naturalistic text, language models struggle with maintaining story coherence — the logical progression of events — and may also become repetitive. Attempts to enhance the coherence of the stories and control the trajectory of events often decompose the generation task into *planning* an outline or sketch, and then *elaborating* on it, e.g., by filling in descriptions, and specific details of each story. Story plans have been previously represented by keywords such as events or phrases (Yao et al., 2019; Xu et al., 2018; Rashkin et al., 2020), a sequence of actions (Goldfarb-Tarrant et al., 2020;

Fan et al., 2019), optionally combined with information on characters and setting (Yang et al., 2022), and control tokens (Lin and Riedl, 2021; Peng et al., 2018; Ippolito et al., 2019; Xu et al., 2020).

Previous work has demonstrated the effectiveness of planning for story generation exclusively in a monolingual setting focusing primarily on English (Fang et al., 2021; Alhussain and Azmi, 2021). We consider whether planning brings advantages to automatic story generation across languages. Specifically, we introduce a new task of *crosslingual* story generation: given a plan in English, generate a coherent narrative in a target language which is consistent with the contents of the plan. The task is challenging for several reasons. Firstly, the plans provide only basic plot information about the story. The model needs to flesh out these plot elements and generate fluent text in the target language, while staying true to the original plan. Moreover, there are differences in story telling traditions amongst cultures and languages (Zipes, 2012), which felicitous stories meant for human readers would need to observe.

We generate stories in several languages leveraging the creative and reasoning capabilities of pre-trained language models (Chowdhery et al., 2022; Wei et al., 2022b) which have recently achieved strong performance on various tasks and in many languages (Wei et al., 2022a; Hao et al., 2022; Arora et al., 2022). PLMs provide a unified modeling framework for our task. Rather than building a different model for language, the PLM is used to generate stories in any target language, without further modification. Following Brown et al. (2020),

we consider a class of hand-designed prompts referred to as "prompting" or "in-context learning". The prompt starts with a free form instruction, followed by a small number of instances exemplifying how the task is solved. In our case, prompts contain plans and their corresponding stories.

We further investigate which plans are best suited for crosslingual story generation. Drawing inspiration from theories of discourse structure (Grosz et al., 1995; Carlson, 1983; Roberts, 2012) and narrative fiction (McKee, 1997), we create plan representations which differ in form and content. As far as the content is concerned, we formulate plans as a list of entities, a sequence of actions, and events revolving around the three-act structure (Field, 2005), a popular story writing technique that divides a narrative into three distinct parts: the setup, the conflict, and the resolution. In terms of form, plans are keywords, prosaic text, and question-answer pairs. To facilitate research in this area, we further create a new dataset containing stories in 31 linguistically diverse languages collated from the Global African Storybook Project.[1]

Our results demonstrate that the three-act structure leads to better stories (across languages) compared to less detailed alternatives. We hypothesize this is due to the inner structure of the plan, i.e., the questions inform the model of important events (i.e., setup, plot, and resolution) and improve the coherence of the story. In addition, our work shows that large language models are capable of generating fluent narratives in multiple languages, while mostly being pretrained on English data. Our contributions can be summarized as follows: (a) we propose a new task of crosslingual story generation with planning and present a new dataset[2] for this task; (b) we conduct a comprehensive study of different plans and prompt formulations for crosslingual story generation; and (c) present results, based on automatic and human assessments, which confirm that plans following the three-act structure expressed as question-answer pairs are more controllable and generate more coherent narratives.

2. Related Work

Story Generation with Planning Plans have been widely used to improve coherence in automated story generation. Several approaches (Yao et al., 2019; Fan et al., 2019; Goldfarb-Tarrant et al., 2020; Fang et al., 2021; Yang et al., 2022) have relied on planning as an intermediate step before generating the full story, with all narrative details. Intermediate plans have taken the form of a sequence of entities and their actions (Yao et al., 2019), outlines based on semantic role labeling (Fan et al., 2019), plot structures rooted in Aristotelian philosophy (Goldfarb-Tarrant et al., 2020), and more elaborate descriptions including details about the setting of the story, its characters, and main plot points (Yang et al., 2022). These efforts have consistently demonstrated that generating full stories based on structured plans improves their coherence and overall quality. We build on this work in two ways: a) we study plan-based story generation in a crosslingual context, which, to our knowledge, has not been done so far; and b) we systematically study which type of plan leads to better stories in this crosslingual setting. We simplify the generation problem in that plans are not learned or predicted (e.g., Goldfarb-Tarrant et al. 2020; Yang et al. 2022), instead they are provided to the model to generate from. Future work could explore how to generate plans in addition to stories given an initial specification (Yang et al., 2022).

Work on multilingual story generation has made little progress due to scarcity of resources (Hou et al., 2019). MTG (Chen et al., 2021) is a new multilingual benchmark covering several subtasks one of which is story generation. Specifically, they translate ROCStories (Mostafazadeh et al., 2016), a widely used dataset for testing performance on the Story Cloze task, into four languages (de, fr, es, zh). We create the first crosslingual dataset which allows to evaluate full stories (as opposed to story completions) in a crosslingual setting.

Prompting Prompting aims to make better use of the knowledge encapsulated in pretrained language models to solve various types of downstream tasks by simply conditioning on a few examples (few-shot) or instructions describing the task at hand (zero-shot). Existing work (Brown et al., 2020; Schick and Schütze, 2021, *inter alia*) has shown that prompting can lead to strong performance in a wide range of tasks including question answering (Chowdhery et al., 2022; Agrawal et al., 2022), and open-ended natural language generation (Tang et al., 2022). Prompting in multilingual settings has achieved good performance using English prompts and target language exemplars (Winata et al., 2021; Lin et al., 2021; Shi et al., 2022a). Polyglot prompting (Fu et al., 2022) aims to learn a unified semantic space for different languages based on multilingual prompt engineering.

Our work employs prompting in a crosslingual setting. We adopt a unified modeling approach, we generate stories in different languages with same model, by changing the language id and the story examples presented to the model, while English

---

[1]Global ASP aims to translate African stories into all of the world's languages.

[2]We will make the dataset available at URL to foster future work on crosslingual story planning and generation.

Amdo Tibetan (28)	Haitian Creole (28)	Polish (28)
Arabic (28)	Hindi (28)	Punjabi (28)
Bengali (28)	Hungarian (28)	Romanian (28)
Chinese (28)	Italian (28)	Russian (10)
Danish (28)	Japanese (28)	Spanish (28)
Dutch (8)	Khams Tibetan (28)	Swedish (28)
Esperanto (28)	Korean (28)	Tibetan (28)
German (28)	Kurdish (28)	Turkish (28)
Greek (28)	Pashto (28)	Urdu (28)
Gujarati (4)	Persian (28)	Vietnamese (28)
		Yue Chinese (28)

Table 1: Number of stories (within parentheses) translated from English in our ASPEN dataset.

Languages	Length	Num Sents	Num Stories
en	482.36	33.61	28
de	540.82	29.32	28
ru	490.80	27.80	10
it	580.89	33.25	28

Table 2: Statistics of ASPEN for the focus languages, German, Russian and Italian. Length is measured in SentencePiece tokens (based on mT5-large).

plans remain fixed.

3. The ASPEN Benchmark for Crosslingual Story Generation

Task Description Given a plan in English, our task is to generate a coherent narrative in a target language which is consistent with the contents of the plan. This renders our setting strictly crosslingual, i.e., we assume models do not observe any target language text in the plan. We expect plans to serve as structured information about story content, leaving room for generating stories creatively in different languages.

Dataset Creation Our dataset, ASPEN, is based on the Global **A**frican **S**tory **P**roject which we collate to serve the crosslingual g**EN**eration task sketched above. Global ASP consists of over 40 stories and their translations into 55 languages. In creating ASPEN, we only include stories with an English version so that English plans can be created, and (in experiments) focus on target stories in Russian, Italian, and German. However, we are releasing stories in 31 languages together with the associated plans (see Table 1 for the distribution of stories per language).

Global ASP stories are based on illustrated children's books, and as such a few of them do not have a narrative plot, they mostly contain short demonstrative sentences (e.g., "Here is a zebra. Here is an elephant."). In constructing ASPEN, we only include stories whose average sentence length (in the English version) is over 6 tokens.[3] Table 2 provides statistics of the dataset. We measure story length in SentencePieces and the number of sentences is calculated using a sentence splitter based on NLTK (Bird et al., 2009)). Plans are created manually based on English stories, while stories in the target languages (ru, de, it) are their parallel translations. We provide details on the content

and structure of the plans in Section 4. Note that Russian comprises only 10 suitable stories. We discuss in Section 5 how we create training/test partitions for our few-shot experiments. The dataset is available at ASPEN URL.

Given its small size, ASPEN is intended primarily as an evaluation corpus. However, it contains several languages (see Table 1), even though to make evaluation tractable, we concentrate on three.

4. Problem Formulation

We next provide an overview of our approach. We follow the few-shot prompting line of research where a PLM is prompted with a few labelled examples and then tasked to complete a test example based on user input.

4.1. Prompt Engineering

Recent work has demonstrated the success of *prompting* general-purpose language models on a wide range of tasks, without the need for modifying any of the model's parameters. Essentially, the model is given an instruction in natural language describing the task at hand (zero-shot setting), optionally followed by demonstrations, i.e., a small number of instances exemplifying how the task is solved (few-shot setting).

Previous work (Shi et al., 2022b) has shown that model output can be fragile and highly dependent on the formulation of the prompt, especially in multilingual settings. Following Reynolds and Mc-Donell (2021), we use "anthropomorphic" prompts, i.e., prompts which would be self-explanatory when shown to a person. Our prompt is illustrated in Figure 1. As can be seen, a demonstration includes the plan of a story (in English), the name of the target language, and the target story verbalizing the plan. It is followed by a test prefix which generates a story based on the provided plan and target language id. In experiments, we used three plan-story demonstrations, however, we only show one in Figure 1 for the sake of brevity.

An important variable in the prompt is the story prefix. This phrase directly precedes the example

---

[3]By tokens here we refer to SentencePieces obtained from the mT5-large pretrained Tokenizer.

Three-act structure	*What is the setup?* Goat, Dog and Cow took a taxi together. *What is the main conflict?* They needed to pay for the ride. *What solution do the characters find?* Cow paid her fare. Dog paid a bit extra. Goat ran away without paying anything. *What does the story end with?* Dow runs towards the car to find the driver who owes him change. Cow Crosses the road calmly because she pays her fare. Goat runs away from the car ecause she is afraid to be arrested.
Language	German
Story	Ziege, Hund und Kuh waren gute Freunde. Eines Tages machten sie eine Fahrt mit einem Taxi. Als sie das Ende ihrer Fahrt erreichten, forderte der Fahrer sie auf, ihren Fahrpreis zu bezahlen. Kuh zahlte ihren Fahrpreis. Hund zahlte etwas mehr, weil er nicht das richtige Geld hatte. Der Fahrer wollte Hund gerade sein Wechselgeld geben, als Ziege wegleif, ohne etwas zu bezahlen. Der Fahrer war sehr verärgert. Er fuhr davon, ohne Hund sein Wechselgeld zu geben. Deshalb rennt Hund auch heute noch zu einem Auto, um hineinzuspähen und den Fahrer zu finden, der ihm sein Wechselgeld schuldet. Ziege rennt vor dem Geräusch eines Autos davon. Sie befürchtet, verhaftet zu werden, weil sie ihren Fahrpreis nicht bezahlt hat. Und Kuh stört es nicht, wenn ein Auto kommt. Cow nimmt sich Zeit, die Straße zu überqueren, weil sie weiß, dass sie ihren Fahrpreis vollständig bezahlt hat.
Three-act structure	*What is the setup* Andiswa wanted to play soccer. *What is the main conflict?* The coach told Andiswa that only boys are allowed to play soccer. *What solution does the character find?* During a big mtch the best soccer player was ill. Andiswa asked the coach to play instead of him. The coach allowed her to play. *What does the story end with?* Andiswa scored a goal. The crowd was joyful. Since then the girls are allowed to play soccer.
Language	German
Story	

Figure 1: Example prompt for crosslingual story generation. First, a few training examples are demonstrated to the model (in green rectangle) and then the model has to complete a prefix for a test example (in yellow rectangle).

Story Prefixes

1. Story:
2. A native ⟨tgt_language⟩ speaker would write the story as:
3. Die Geschichte: (⟨"Story:" in tgt_language⟩):
4. ⟨tgt_language⟩ story:

Table 3: Story Prefixes used within different prompts. ⟨tgt_language⟩ is a paceholder for German, Russian, or Italian in our experiments. "Die Geschichte" is the translation of "Story" in German (for Russian it would be историю, and "Storia" for Italian).

story shown to model and is at the very start of the instance the model is supposed to complete (at generation time). As it directly precedes the generated story, this part of the prompt greatly influences the output. We experimented with the four story prefixes shown in Table 3.

4.2. Plan Representation

In our crosslingual setting, we rely on plans to convey the content of the story to the model. In experiments, we use manually created plans which vary in terms of their form and the level of plot detail supplied to the model. We describe these below and provide examples in Figure 2.

Story Completion Much previous work has focused on generating a continuation or ending for an incomplete story (Mostafazadeh et al., 2016; Fan et al., 2018; Wang and Wan, 2019; Mori et al., 2022). We adapted story completion to our crosslingual setting, by prompting the model with the first two sentences of the English story. Although the beginning of the story is not a plan as such, we assume it contains enough detail about the characters and the setting of the story (see Figure 2). Our intuition is that the model will learn that the task involves translating the incomplete story to the target

language and then generating a continuation.

Entities Content planning strategies based on entities have been proven effective in a variety of tasks beyond story generation, including summarization (Narayan et al., 2021; Liu and Chen, 2021) and data-to-text generation (Puduppully and Lapata, 2021). Entities also play a pivotal role in various theories of discourse which posit that coherence is achieved in view of the way discourse entities are introduced and discussed (Grosz et al., 1995). Our entity-based plans are a list of characters, places, and objects in the story (see Figure 2).

Plot Outline Much research on story generation (Goldfarb-Tarrant et al., 2020; Rashkin et al., 2020; Fan et al., 2019, *inter alia*) has resorted to plot outlines as a means of instilling generation models with knowledge about events and their logical progression. Stories in our dataset are accompanied by short, high-level summaries which give an overview of the plot without going into detail. We used these summaries as a proxy for plot outlines (see the example in Figure 2).

Three-act Structure The three-act structure is a model used in narrative fiction that divides a story into three parts (acts), often called the Setup, the Confrontation, and the Resolution. (McKee, 1997; Field, 2005). The setup describes the overall circumstances of the story, e.g., main characters and their life. The confrontation describes the conflict which the main character needs to overcome. Resolution describes the steps the protagonist takes to resolve the conflict and their outcome. We devised plans based on the three-act structure under the hypothesis that stories attempt to answer the following questions: What is the setup?, b) What is the main conflict?, c) What solution does the character find?, and d) What does the story end with? All

Story Completion	Goat, Dog, and Cow were great friends. One day they went on a journey in a taxi.
Entities	Goat, dog, Cow, friends, they, they, the driver, them, their ares, Cow, Dog, he, The driver, Dog, Goat, The driver, He, Dog, Dog, the driver, him, Goat, She, she, Cow, Cow, she, she
Plot Outline	Goat, Dog and Cow are good friends. They take a trip together on a taxi, but when the time comes to pay the driver one of the friends does something surprising.
Three-act Structure	*What is the setup?* Goat, Dog and Cow took a taxi together. *What is the main conflict?* They needed to pay for the ride. *What solution do the characters find?* Cow paid her fare. Dog paid a bit extra. Goat ran away without paying anything. *What does the story end with?* Dog runs towards the car to find the driver who owes him change. Cow crosses the road calmly because she paid her fare. Goat runs away from the car because she is afraid to be arrested.

Figure 2: Story plan examples. Questions in italics correspond to main events in three-act structure.

plans have the same questions, however, the answers vary depending on the content of individual stories (see Figures 1 and 2).

5. Experimental Setup

5.1. Models

As mentioned earlier, we employ a single model for all story generation experiments. Specifically, we use PaLM (Chowdhery et al., 2022) [4] which was mostly trained on English (80% of training data). However, PaLMs have recently demonstrated impressive results in multilingual settings, especially when prompted with several training examples (Winata et al., 2021; Chowdhery et al., 2022; Shi et al., 2022b). Throughout this paper, PaLM was called with temperature $\tau = 0.7$, allowing for some randomness in the output, with a beam width of $w = 10$. All experiments used the prompts described in Section 4, with three (randomly sampled) plan-story examples per language. We use the same three examples for all languages and treat the remaining stories in ASPEN as test data. Plans varied along the dimensions introduced in Section 4.2.

As a baseline, we used another strong state-of-the-art multilingual model, namely mT5 (Xue et al., 2021a). mT5 was pretrained on data in multiple languages, making it especially suited to our crosslingual generation task. We finetuned mT5-XL (Xue et al., 2021b) on three plan-story examples per language [5] (same as those seen by the PLM) for 500 steps with a batch size of 8 and a learning rate of 0.0001. We selected the last checkpoint for each language. As an upper bound we further translated the English stories into the target language using the public Google Translate API (Wu et al., 2016).

5.2. Evaluation

Story generation is a creative task, the model is allowed to generate new content which is not included in the input (Deng et al., 2021). In our case, not only is the model creating new text, but also the story is expected to be in a different target language.

These constraints make evaluation challenging, the generated text needs to be fluent in the target language (no repetitions/grammar errors), while the story needs to be overall coherent, and faithful to the English plan. We evaluate these aspects both automatically and in a judgment elicitation study.

Automatic Evaluation We evaluated various aspects of fluency following the automatic metrics introduced in Goldfarb-Tarrant et al. (2020). Specifically, we use *vocabulary-to-token ratio* to measure the extent to which the vocabulary of the generated stories is repetitive (higher is better). We further measure *intra-story repetition* as a fluency metric using the proportion of trigrams which are repeated *within* a story (lower is better). Finally, we also compute *inter-story repetition* as a diversity metric to quantify whether the model has learnt to generate only one "kind" of text irrespective of the input. We measure the proportion of trigrams repeated *between* stories (lower is better).

We also use MAUVE (Pillutla et al., 2021) to measure the naturalness of the generated stories. MAUVE is a recently introduced automatic metric for open-ended generation which has high correlation with human judgements. It computes the similarity of the distribution of human-written text and machine-generated text, while being sensitive to to generation length, different decoding algorithms, and model size. Machine-generated distributions in MAUVE are computed with GPT-2 (Radford et al., 2019). We use gold reference stories in the target language as our set of human written texts.

Finally, we also evaluate the similarity of model output against reference stories in ASPEN. As the stories are parallel in the dataset, we expect that the closer the model follows the plan based on the English story, the more similar the resulting story will be to the golden stories. Notably, the reference stories are manual native spaker translations of the stories to the target languages. We chose these as golden stories to compare against non-creative, translation only baseline. We use SentencePieceR-OUGE (Vu et al., 2022), a ROUGE-inspired (Lin, 2004) metric which uses language-independent SentencePiece (Kudo and Richardson, 2018) tokenization.

	DE			RU			IT			AVG		
Models	VocTok ↑	Inter ↓	Intra ↓	VocTok ↑	Inter ↓	Intra ↓	VocTok ↑	Inter ↓	Intra ↓	VocTok ↑	Inter ↓	Intra ↓
Entities	0.39	22.63	1.75	0.34	48.36	3.15	0.37	26.62	1.35	0.37	32.53	**2.09**
Story Completion	0.43	33.53	4.04	0.38	51.08	**0.72**	0.61	23.09	3.81	0.47	35.90	2.86
Plot Outline	0.41	38.06	10.03	**0.65**	22.09	2.64	0.47	21.18	**1.03**	0.51	27.11	4.57
3Act Structure	**0.57**	**18.85**	**1.68**	0.58	**14.66**	3.51	**0.48**	**16.00**	1.52	**0.54**	**16.50**	2.24
mT5	0.58	91.17	0.00	0.73	68.34	0.42	0.68	87.77	0.00	0.66	82.43	0.14
Google Translate	0.57	6.85	0.74	0.70	4.08	0.21	0.58	7.23	0.42	0.62	6.05	0.46
Reference	0.58	8.02	0.50	0.65	5.16	2.22	0.57	7.61	0.52	0.60	6.93	1.08

(PaLM: Entities, Story Completion, Plot Outline, 3Act Structure)

Table 4: Diversity and repetitiveness metrics for PaLM with different plan variants and comparison models.

Models	DE	RU	IT	AVG
Entities	0.95	0.91	0.99	**0.95**
Story Completion	**0.99**	0.66	0.99	0.88
Plot Outline	0.98	0.20	0.99	0.72
3Act Structure	**0.99**	0.66	0.99	0.89
mT5	0.37	**0.99**	0.86	0.74
Google Translate	0.99	0.20	**1.00**	0.73

(PaLM: Entities, Story Completion, Plot Outline, 3Act Structure)

Table 5: MAUVE for PaLM with different plan variants and comparison models.

Models	EN	DE	RU	IT	AVG
Entities	21.52	20.19	20.32	19.92	20.14
Story Completion	20.07	19.71	14.17	20.24	18.04
Plot Outline	19.38	18.10	18.46	18.60	18.39
3Act Structure	23.52	**22.61**	**21.16**	**24.34**	**22.70**
mT5	N/A	15.84	12.70	15.25	14.60
Google Translate	N/A	73.98	68.31	68.88	70.39

(PaLM: Entities, Story Completion, Plot Outline, 3Act Structure)

Table 6: SentencePiece-ROUGE between generated and reference story for PaLM with different plan variants and comparison models. EN results are provided for reference and excluded from AVG.

Human Evaluation We also carry out human evaluation following a simplified version of the annotation protocol proposed in Chhun et al. (2022). Specifically, crowdworkers are presented with the English prompt given to the model, the human-authored story in the target language, followed by the machine generated story, also in the target language. We provided reference stories so that participants could better calibrate their judgment (Karpinska et al., 2021).

Participants are asked to rate the stories along the following dimensions: (a) *Relevance* measures whether the story matches its prompt; (b) *Fluency* measures the quality of the text including grammatical errors and repetitions; (c) *Coherence* measures whether the story's plot makes logical sense; and (d) *Engagement* measures the extent to which the reader engaged with the story. Each story was evaluated by three workers on the four criteria using a 3-point Likert scale where 1 is worst and 3 is best. Annotators were native speakers of the target language, and fluent in English. We evaluate the output of our model with the three plan configurations, mT5, and Google Translate (without a plan). Overall, we elicited ratings for 285 stories (35 in Russian, 125 in Italian, and 125 in German). We collect ratings from three different annotators for each data point. Figure 3 in Appendix A presents our detailed instructions.

6. Results

6.1. Automatic Evaluation

Tables 4–6 summarize our results using automatic evaluation metrics. We present results with different instantiations of our model according to the plans presented in Section 4.2 and the prefix "<tgt_language> story:" (see Table 3) which performed overall best. We describe results with alternative story prefixes in Appendix B. We also compare to an mT5 model fine-tuned in a similar few-shot setting, and Google Translate as an upper bound. Wherever possible we also report metrics on the gold reference translations.

How Diverse are the Generated Stories? In Table 4 we report results with vocabulary to token ratio (VocTok), Inter- and Intra-story repetition for each target language, and on average. As can be seen, stories following the three-act structure have the most diverse vocabulary (see VocToK column), they also tend to be overall fluent (see Intra column). Inter-story repetition measures whether the constructions used are repetitive across stories. Ideally, we would like to avoid cases where the model generates the same story no matter the plan. Table 4 shows that the three-act plan leads to more diverse texts. We hypothesize it forces the model to focus on main events and their ordering while other plan formulations exercise less control on the content and structure of the story. The mT5 model performs poorly, especially with respect to the Inter-story repetition metric, which indicates that it generates the same story irrespective of the plan. The Google Translate upper bound performs quite well, approaching human parity (see row Reference in Table 4).

Do Stories Resemble Human Writing? We now examine whether the generated stories are similar to human-written texts. Table 5 quantifies the naturalness of the stories using MAUVE (Pil-

Models	DE	RU	IT	AVG
PaLM Entities	551.50	136.39	656.15	448.01
Story Completion	505.50	**94.23**	453.15	350.96
Plot Outline	529.52	124.65	541.27	398.30
3Act Structure	**386.62**	99.81	431.31	**305.91**
mT5	494.96	260.00	**355.08**	370.01
Google Translate	557.08	461.71	598.12	538.97
Reference	548.08	453.29	585.32	528.90

Table 7: Average story length (measured in SentencePiece tokens) for PaLM and comparison models.

Models	AVG
PaLM Entities	66.66
Story Completion	42.31
Plot Outline	59.64
3Act Structure	38.58
mT5	77.19
Google Translate	66.66
Reference	68.42

Table 8: Percentage of stories containing direct speech for PaLM and comparison models.

Models	Relevance	Fluency	Coherence	Engagement
PaLM Entities	0.37°	0.51°	0.46°	0.33°
Plot Outline	0.36°	0.55°	0.50°	0.31°
3Act Structure	**0.59**	**0.61°***	**0.69***	**0.46°**
mT5	0.00	**0.82**	**0.69***	0.42°
Google Translate	0.99	0.68*	0.92	0.60

Table 9: Average human evaluation results across languages for PLM with different plan variants and comparison models. Best results in upper block are **boldfaced**. Systems in each column are marked with same symbols when differences between them are not statistically significant; unmarked pairwise differences are significant (using a one-way ANOVA with post-hoc Tukey HSD tests; $p < 0.01$).

lutla et al., 2021). In general, we observe that the similarity of model output against human-written text is high, particularly for German and Italian. Plans based on entities impose least constraints on the output, and therefore lead to most natural text. Plans following the three-act structure are second best according to MAUVE. We also see that MAUVE penalizes Google Translate, we suspect it is prone to translationese which renders the stories less natural. The same is true for the mT5 model.

Are Machine and Reference Stories Similar? In Table 6 we evaluate similarity between automatically generated and "gold" reference stories using SentencePiece-ROUGE. This allows us to simultaneously measure whether the events in the gold story are mentioned in the generated story and whether the language usage is similar between reference stories and automatically generated ones. Across all languages, we observe that stories based on the three-act structure are more similar to the reference. This is not surprising given this plan is most detailed; the story completion and plot outline plans cover most of the content of the story but lack details. The model based on entity plans performs quite well even though the plan is not very elaborate. These stories tend to be fairly long (see Table 7) and thus favored by recall-oriented ROUGE. mT5 struggles to generate stories which follow the plan is penalized by ROUGE. In general, all models have a long way to go compared to the upper bound (see Google Translate in the table).

Do Stories Have Different Length? The plans can be viewed as constraints on open-ended generation. In other words, we expect that with plans containing more information, the model will have a clearer representation of the content of the story and will stop generating once all of the planned content is covered. To test this hypothesis, we compare the average length of stories generated based on different plans (we measure length in SentencePieces based on the vocabulary pretrained for mT5-large). The results in Table 7 corroborate our hypothesis. Plans based on three-acts contain most information about events in the story and their order, resulting in the shortest narratives. In contrast, entity-based plans are least constraining, they do not contain information about relations between characters or the order of events and as a result the model has free reign to fill in missing content and ends up generating longer stories. All models, including mT5, generate shorter stories compared to the Reference and Google Translate.

Do Stories Contain Direct Speech? The plans used in our experiments are prosaic, containing no examples of direct speech. However, dialogue is a common device for rendering narratives more engaging, moving the story forward, and allowing characters to engage in conflict (Scott Bell, 2014). We evaluate the extent to which a model is creative by computing the percentage of stories which include direct speech. We assume direct speech is marked with pairs of quotation marks. The results in Table 8 show that plans based on entities and three-acts display the most and least direct speech, respectively. This further confirms our observation that with less constraining plans the model is being more creative with its output, but at the cost of coherence (as we discuss below). Interestingly, the percentage of stories with direct speech generated by mT5 is seemingly high; this is because the model generates the same story irrespective of the input plan in most cases. As this story contains an example of direct speech, this percentage is high.

Entities
Una volta, c'era una mucca, un cane e una capra che andarono a visitare un altro villaggio. Decisero di prendere un autobus. Il cane era un po' preoccupato, ma la capra era molto felice e disse al cane: "Non preoccuparti! La gente sarà gentile con noi." Il cane non era sicuro, ma la capra era determinata. Così tutti e tre i loro amici salirono sull'autobus. Quando il conducente dell'autobus li vide salire, cominciò a ridere. "Siete voi tre amici che vi siete persi?" Disse il conducente. "Dove state andando?" La capra, che era più intelligente dei suoi amici, rispose: "Stiamo andando a visitare un altro villaggio." Il conducente dell'autobus si mise a ridere di nuovo. "Bene, amici! Non ho mai visto animali sull'autobus! Siete molto divertenti!" Disse. Il cane non era sicuro. La capra rispose: "Abbiamo pagato abbastanza soldi per prendere l'autobus."

3Act Structure
Una volta, la mucca, la capra e il cane viaggiarono insieme sul taxi. <...> Chiese la mucca. "Se non pagherai, il conducente del taxi chiamarà la polizia e sarai arrestata.Non puoi semplicemente scappare!" La capra non ascoltò e corse via. Il cane si precipitò verso il taxi per chiedere il resto all'autista. La mucca attraversò la strada con calma, perché aveva pagato il giusto prezzo. La capra uscì correndo dal taxi e scappò via. La capra non pagò il taxi. La capra fuggì. La mucca attraversò la strada con calma.

Table 10: Abridged stories generated in Italian from different plans (story ID 4: "Goat, Dog and Cow"). Highlighted excerpts correspond to the events and/or characters mentioned in the input plan.

6.2. Human Evaluation

Our judgment elicitation study is summarized in Table 9. Our results are averaged across languages, we present ratings for each language in Appendix A). Human evaluation confirms our automatic results. The model based on the three-act structure manages to generate stories which 'stick' to the plan (see Relevance column), whereas models based on entities and plot outlines are less faithful (the difference between the these two and the three acts is statistically significant), while mT5 largely ignores the plan. All PLM variants are equally fluent (differences among them are not statistically significant). The three-acts model is as fluent as Google Translate, however, the most fluent model is mT5 which is good at generating text but not necessarily the story we want.

Human raters find the stories based on the three-acts as coherent as those obtained from mT5 and both of these (significantly) less coherent than Google Translate. The stories obtained from the three acts are more engaging compared to entities and plot outlines, but less so compared to the Google Translate upper bound (which is significantly better against all other models). We show examples of stories our raters saw in Table 10. We present analysis of system output in Appendix C.

7. Conclusions

In this work we considered the problem of automatically generating stories in multiple languages by prompting pretrained language models with different plans. We investigated which plan formulation is better suited to our crosslingual generation task and empirically demonstrated that plans following the three-act structure generate more coherent narratives. To facilitate research in this area, we further collated ASPEN, a new dataset with multilingual stories and plan annotations.

In the future, we would like to explore further the potential of formulating plans as question-answer pairs, to generate stories that are both creative and controllable. For example, we could generate more detailed questions with very short answers or no answers at all. It would also be interesting to generate the plan and the story iteratively allowing human writers to intervene, and directly steer the generation process. Further, our experiments have not explored how the number of example stories provided to the model at inference time influences the output in a creative task such as story generation. We leave it as a potential future research.

In this paper we mainly focused on three Indo-European languages, namely Russian, German, and Italian. Although the proposed methods are in principle language agnostic, the PLM's performance might vary depending on the number and amount of languages it has been trained on. Another limitation concerns the use of the large language model itself which is feasible only with large computational resources. Benchmarking LLMs on ASPEN is out of scope of this paper and we leave it as future work, together with other cross-lingual transfer methods, e.g., translation-based. Finally, throughout this paper, we have assumed that the plans are provided to the PLM, e.g., in an interactive setting where a user thinks of a story sketch, and the system fleshes out the details. In the future, it would be interesting to consider generating plan candidates as well as stories.

References

Priyanka Agrawal, Chris Alberti, Fantine Huot, Joshua Maynez, Ji Ma, Sebastian Ruder, Kuzman Ganchev, Dipanjan Das, and Mirella Lapata. 2022. Qameleon: Multilingual qa with only 5 examples.

Arwa I Alhussain and Aqil M Azmi. 2021. Automatic story generation: a survey of approaches. *ACM Computing Surveys (CSUR)*, 54(5):1–38.

Simran Arora, Avanika Narayan, Mayee F Chen, Laurel J Orr, Neel Guha, Kush Bhatia, Ines Chami, Frederic Sala, and Christopher Ré.

2022. Ask me anything: A simple strategy for prompting language models. *arXiv preprint arXiv:2210.02441*.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

L Carlson. 1983. *Dialogue Games: An Approach to Discourse Analysis*. Riedel, Dordrecht.

Marc Cavazza, Olivier Martin, Fred Charles, Steven J Mead, and Xavier Marichal. 2003. Interactingwith virtual agents in mixed reality interactive storytelling. In *International Workshop on Intelligent Virtual Agents*, pages 231–235. Springer.

Yiran Chen, Zhenqiao Song, Xianze Wu, Danqing Wang, Jingjing Xu, Jiaze Chen, Hao Zhou, and Lei Li. 2021. Mtg: A benchmarking suite for multilingual text generation. *arXiv preprint arXiv:2108.07140*.

Cyril Chhun, Pierre Colombo, Fabian M. Suchanek, and Chloé Clavel. 2022. Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5794–5836, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Mingkai Deng, Bowen Tan, Zhengzhong Liu, Eric Xing, and Zhiting Hu. 2021. Compression, transduction, and creation: A unified framework for evaluating natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7580–7605.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898.

Angela Fan, Mike Lewis, and Yann Dauphin. 2019. Strategies for structuring story generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2650–2660.

Le Fang, Tao Zeng, Chaochun Liu, Liefeng Bo, Wen Dong, and Changyou Chen. 2021. Outline to story: Fine-grained controllable story generation from cascaded events. *arXiv preprint arXiv:2101.00822*.

Syd Field. 2005. *Screenplay: The foundations of screenwriting*. Delta.

Jinlan Fu, See-Kiong Ng, and Pengfei Liu. 2022. Polyglot prompt: Multilingual multitask promptraining. *arXiv preprint arXiv:2204.14264*.

Seraphina Goldfarb-Tarrant, Tuhin Chakrabarty, Ralph Weischedel, and Nanyun Peng. 2020. Content planning for neural story generation with aristotelian rescoring. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4319–4338.

Barbara Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Yaru Hao, Haoyu Song, Li Dong, Shaohan Huang, Zewen Chi, Wenhui Wang, Shuming Ma, and Furu Wei. 2022. Language models are general-purpose interfaces. *arXiv preprint arXiv:2206.06336*.

Chenglong Hou, Chensong Zhou, Kun Zhou, Jinan Sun, and Sisi Xuanyuan. 2019. A survey of deep learning applied to story generation. In *International Conference on Smart Computing and Communication*, pages 1–10. Springer.

Daphne Ippolito, David Grangier, Chris Callison-Burch, and Douglas Eck. 2019. Unsupervised hierarchical story infilling. In *Proceedings of the First Workshop on Narrative Understanding*, pages 37–43, Minneapolis, Minnesota. Association for Computational Linguistics.

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The perils of using Mechanical Turk to evaluate open-ended text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1265–1285, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text

processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Michael Lebowitz. 1987. Planning stories. In *Proceedings of the 9th annual conference of the cognitive science society*, pages 234–242.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona T. Diab, Veselin Stoyanov, and Xian Li. 2021. Few-shot learning with multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhiyu Lin and Mark Riedl. 2021. Plug-and-blend: A framework for controllable story generation with blended control codes. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 62–71, Virtual. Association for Computational Linguistics.

Hugo Liu and Push Singh. 2002. Using commonsense reasoning to generate stories. In *Proceedings of the 18th National Conference on Artificial Intelligence*, pages 957–958, Edmonton, Alberta.

Zhengyuan Liu and Nancy Chen. 2021. Controllable neural dialogue summarization with personal named entity planning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 92–106, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Robert McKee. 1997. *Story: Substance, Structure, Style, and the Principles of Screenwriting*. ReganBooks, New York, NY.

James Meehan. 1977. An interactive program that writes stories. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence*, pages 91–98, Cambridge, Massachusetts.

Yusuke Mori, Hiroaki Yamane, Ryohei Shimizu, and Tatsuya Harada. 2022. Plug-and-play controller for story completion: A pilot study toward emotion-aware story writing assistance. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*, pages 46–57, Dublin, Ireland. Association for Computational Linguistics.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. 2021. Planning with learned entity prompts for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 9:1475–1492.

Federico Peinado and Pablo Gervás. 2005. Creativity issues in plot generation. In *Workshop on Computational Creativity*, Working Notes, 19th International Joint Conference on A, pages 45–52.

Nanyun Peng, Marjan Ghazvininejad, Jonathan May, and Kevin Knight. 2018. Towards controllable story generation. In *Proceedings of the First Workshop on Storytelling*, pages 43–49, New Orleans, Louisiana. Association for Computational Linguistics.

Rafael Pérez y Pérez and Mike Sharples. 2001. Mexica: A computer model of a cognitive account of creative writing. *Journal of Experimental and Theoretical Artificial Intelligence*, 13(2):119–139.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34:4816–4828.

Ratish Puduppully and Mirella Lapata. 2021. Data-to-text generation with macro planning. *Transactions of the Association for Computational Linguistics*, 9:510–527.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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

Hannah Rashkin, Asli Celikyilmaz, Yejin Choi, and Jianfeng Gao. 2020. PlotMachines: Outline-conditioned generation with dynamic plot state tracking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4274–4295, Online. Association for Computational Linguistics.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Mark O Riedl and Robert Michael Young. 2010. Narrative planning: Balancing plot and characte. *Journal of Artificial Intelligence Research*, 39:217–268.

Craige Roberts. 2012. Information structure in discourse: Towards an integrated formal theory of pragmat ics. *Semantics and Pragmatics*, 5(6):1–69.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

James Scott Bell. 2014. *How to Write Dazzling Dialogue: The Fastest Way to Improve Any Manuscript*. Compendium Press, London, UK.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022a. Language models are multilingual chain-of-thought reasoners.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022b. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2022. Context-tuning: Learning contextualized prompts for natural language generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6340–6354, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Scott R. Turner. 1992. *Ministrel: A Computer Model of Creativity and Sotrytelling*. University of California, Los Angeles, California.

Tu Vu, Aditya Barua, Brian Lester, Daniel Cer, Mohit Iyyer, and Noah Constant. 2022. Overcoming catastrophic forgetting in zero-shot cross-lingual generation. *arXiv preprint arXiv:2205.12647*.

Tianming Wang and Xiaojun Wan. 2019. T-CVAE: Transformer-based conditioned variational autoencoder for story completion. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5233–5239. International Joint Conferences on Artificial Intelligence Organization.

Stephen Ware and robert Michael Young. 2010. Modeling narrative conflict to generate interesting stories. In *Proceedings of the AAAI Conference on Artificial Intel- ligence and Interactive Digital Entertainment*, volume 5.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language models are few-shot multilingual learners. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 1–15, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg

Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation.

Jingjing Xu, Xuancheng Ren, Yi Zhang, Qi Zeng, Xiaoyan Cai, and Xu Sun. 2018. A skeleton-based model for promoting coherence among sentences in narrative story generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4306–4315, Brussels, Belgium. Association for Computational Linguistics.

Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Raul Puri, Pascale Fung, Anima Anandkumar, and Bryan Catanzaro. 2020. MEGATRON-CNTRL: Controllable story generation with external knowledge using large-scale language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2831–2845, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021a. 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.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021b. 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.

Kevin Yang, Nanyun Peng, Yuandong Tian, and Dan Klein. 2022. Re3: Generating longer stories with recursive reprompting and revision. *arXiv preprint arXiv:2210.06774*.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, pages 7378–7385.

Jack Zipes. 2012. *The Irresistible Fairy Tale: The Cultural and Social History of a Genre*. Princeton University Press, Princeton, NJ.

A. Human Evaluation

Figure 3 presents the instructions shown to crowd-workers during our human evaluation study. To recruit our participants, we screened their language skills to determine whether they're native speakers, their education level and country of residence as well as origin. There were a total of 18 raters, of which 38.89% are from Germany, 38.89% from Ukraine and the rest were from Italy. 61.11% of them holds a master degree, 27.78% of them holds a High school degree or equivalent and 11.11% holds a Bachelor's degree. All our annotators are paid adequately by our suppliers adhering to the supplier code of conduct.

A detailed break down of our human results is provided in Table 11.

B. Results with Alternative Story Prefixes

Table 12–14 present results for our model when using different story prefixes in the prompt.

C. Qualitative Evaluation

Table 15 shows four abridged stories generated in Italian based on different plans. We observe that all of them have the same *main characters* as prescribed in their respective plans (i.e., the goat, the dog and the cow). A secondary character, the taxi driver, is mentioned in stories based on entities and three-acts but is absent when the model is given a plot outline or asked to perform story completion. We hypothesise this is because the taxi driver is not explicitly mentioned in the latter two plan instantiations. There are no new characters added to the story, except in the case of plot outline where two novel characters are present, namely a sheep and a chicken, who the main characters address.

All stories convey the correct *setup* (i.e., travelling between places), However, the story based on three-acts conveys more detail, while others make some mistakes, for instance, the characters use the wrong means of transportation. Most stories have a consistent *flow* of events and final resolution with the exception of the story based on the plot outline which is repetitive (the main character repeatedly asks the same questions). This indicates that coarse outlines might yield less coherent stories, which are less likely to resolve into an ending. No matter what plan is used, the stories are consistently grammatical. We further observed that the stories rarely use pronouns, which are a diagnostic of locally coherent discourse (Grosz et al., 1995).

Our plans contain information which can be used in different parts of the story. The highlights in Table 15 show spans in the story which are directly copied from the plan. As can be seen, entities are interspersed throughout (both as subjects and objects) while story completion is most useful in guiding the beginning of the story (the remainder is being generated creatively). The plot outline is also less constraining leading to more creative but less coherent stories. The PaLM follows the three-act plan closely, drawing information throughout.

Task: Please read the prompt, the human-written story and the machine-generated story, carefully. We will then ask you to rate the machine-generated story.
The prompt will always be in English, but the stories will be in one of the following target languages: German, Italian, or Russian. Both the human-written story and the machine-generated story will be presented in the same target language.

Prompt: ⟨List of Entities, or List of question-answer pairs or a small summary⟩ in English.

Human-written Story: ⟨Human written story in a target language⟩

Machine-generated Story: ⟨Machine-generated story in the same target language⟩

Q1: [Relevance – measures how well the story matches its prompt]
1 – The story has no relationship with the prompt at all (it has a lot of extra details not mentioned in the prompt).
2 – The story roughly matches the prompt (but has a few extra details not mentioned in the prompt).
3 – The story matches the prompt exactly.

Q2: [Fluency – measures the quality of the text. Please make sure not to evaluate the plot of the story, but just the language.]
1 – Hard to understand with multiple grammatical errors and/or multiple repetitions.
2 – Possible to understand but there are some mistakes/repetitions.
3 – Easy to understand without any mistakes/disfluencies.

Q3: [Coherence – measures whether the plot of the story makes sense]
1 – The story does not make sense at all. For instance, the setting and/or characters keep changing, and/or there is no understandable plot.
2 – The story mostly makes sense but is incoherent at places.
3 – The story makes sense overall.

Q4: [Engagement – measures how much you engaged with the story]
1 – You mostly found the story boring. There may be one or two things interesting in the story, but no more.
2 – The story was mildly interesting.
3 – The story almost kept you engaged until the end.

Figure 3: Instructions used in our human evaluation.

		DE				RU				IT			
	Models	Relevant	Fluent	Coherent	Engaging	Relevant	Fluent	Coherent	Engaging	Relevant	Fluent	Coherent	Engaging
PaLM	Entities	0.38	0.58	0.51	**0.38**	0.24	0.45	0.45	0.33	0.49	0.51	0.40	0.27
	Plot Outline	0.34	0.53	0.61	0.28	0.21	0.60	0.36	0.33	0.51	0.53	0.54	0.31
	3Act Structure	**0.49**	0.66	0.73	0.37	**0.57**	**0.64**	**0.69**	**0.62**	**0.69**	0.51	**0.64**	**0.39**
	mT5	0.00	**0.87**	**0.86**	0.37	0.00	0.81	0.67	0.52	0.00	**0.79**	0.55	0.37
	Google Translate	0.97	0.83	0.94	0.51	1.00	0.69	0.93	0.69	0.99	0.51	0.89	0.59

Table 11: Human evaluation (per language) results for PaLM with different plan variants and comparison models. Best results in upper block are **boldfaced**.

		DE			RU			IT			AVG		
Plan	Prefix	VocTok ↑	Inter ↓	Intra ↓	VocTok ↑	Inter ↓	Intra ↓	VocTok	Inter ↓	Intra ↓	VocTok ↑	Inter ↓	Intra ↓
Entities	None	0.38	34.46	3.06	0.38	33.49	5.98	0.50	33.40	2.71	0.42	33.78	3.92
Entities	Long	0.46	22.47	1.55	0.36	30.07	2.77	0.39	26.86	1.37	0.40	26.47	1.90
Entities	Target	0.43	24.02	1.10	0.49	13.30	0.28	0.37	25.19	0.74	0.43	20.84	0.71
Entities	Lang	0.39	22.63	1.75	0.34	48.34	3.15	0.37	26.62	1.35	0.37	32.53	2.09
Story Completion	None	0.68	8.68	2.09	0.60	32.92	8.42	0.79	24.8	4.91	0.69	22.13	5.14
Story Completion	Long	0.38	31.37	4.28	0.63	45.06	16.36	0.55	18.56	0.62	0.52	31.66	7.08
Story Completion	Target	0.49	19.33	2.96	0.55	51.85	11.62	0.44	30.57	3.12	0.49	33.92	5.90
Story Completion	Lang	0.43	33.53	4.04	0.38	51.08	0.72	0.61	23.09	3.81	0.47	35.89	2.86
Plot Outline	None	0.64	5.22	0.71	0.56	33.27	6.99	0.68	21.45	1.81	0.63	19.98	3.17
Plot Outline	Long	0.42	27.88	3.08	0.49	12.49	2.72	0.45	29.52	1.75	0.45	23.30	2.52
Plot Outline	Target	0.36	36.60	1.57	0.45	21.55	0.79	0.42	22.86	4.07	0.41	27.00	2.15
Plot Outline	Lang	0.41	38.06	10.03	0.65	22.09	2.64	0.47	21.18	1.03	0.51	27.1	4.57
3Act Structure	None	0.49	20.18	0.69	0.43	25.86	2.14	0.51	13.03	1.44	0.48	19.69	1.42
3Act Structure	Long	0.57	19.07	1.45	0.47	18.70	1.45	0.53	15.06	0.97	0.53	17.61	1.29
3Act Structure	Target	0.46	17.71	1.19	0.55	31.14	0.65	0.49	15.04	0.96	0.50	21.23	0.93
3Act Structure	Lang	0.57	18.85	1.68	0.58	14.66	3.51	0.48	16.00	1.52	0.54	16.50	2.24

Table 12: Diversity and repetitiveness metrics for PaLM with different plans and story prefixes. None is a shorthand for the prefix "Story:", Long abbreviates prefix "A native ⟨tgt_language⟩ speaker would write the story as:", Target is the translation of the prefix "Story:" in the target language, and Lang refers to "⟨tgt_language⟩ story:".

Plan	Prefix	DE	RU	IT	AVG
Entities	None	0	0.66	0.99	0.55
Entities	Long	0.99	0.20	1	0.73
Entities	Target	0.96	0.66	1	0.87
Entities	Lang	0.95	0.91	0.99	0.95
Story Completion	None	0.74	0.97	0.60	0.77
Story Completion	Long	1	0.20	0.99	0.73
Story Completion	Target	1	0.91	0.77	0.89
Story Completion	Lang	0.99	0.66	0.99	0.88
Plot Outline	None	0.95	0.99	0.24	0.72
Plot Outline	Long	0.99	0.91	1	0.97
Plot Outline	Target	1	0.86	1	0.95
Plot Outline	Lang	0.98	0.20	0.99	0.72
3Act Structure	None	1	0.99	1	0.99
3Act Structure	Long	0.89	0.91	0.99	0.93
3Act Structure	Target	0.99	0.66	1	0.88
3Act Structure	Lang	0.99	0.66	0.99	0.89

Table 13: MAUVE for PaLM with different plan variants and story prefixes. None is a shorthand for the prefix "Story:", Long abbreviates prefix "A native ⟨tgt_language⟩ speaker would write the story as:", Target is the translation of the prefix "Story:" in the target language, and Lang refers to "⟨tgt_language⟩ story:".

Plan	Prefix	DE	RU	IT	AVG
Entities	None	13.21	12.18	13.41	12.93
Entities	Long	20.12	18.87	23.20	20.73
Entities	Target	20.44	17.16	21.75	19.78
Entities	Lang	20.19	20.32	19.92	20.14
Story Completion	None	10.25	17.96	7.96	12.05
Story Completion	Long	20.96	16.94	19.66	19.19
Story Completion	Target	20.47	18.52	19.06	19.35
Story Completion	Lang	19.71	14.17	20.24	18.04
Plot Outline	None	10.93	15.19	7.298	11.14
Plot Outline	Long	17.78	15.94	17.64	17.12
Plot Outline	Target	15.53	16.90	16.81	16.41
Plot Outline	Lang	18.10	18.46	18.60	18.39
3Act Structure	None	22.70	19.80	22.36	21.62
3Act Structure	Long	22.09	21.15	23.16	22.13
3Act Structure	Target	20.99	20.14	24.63	21.92
3Act Structure	Lang	22.61	21.16	24.34	22.70

Table 14: SentencePiece-ROUGE between generated and reference story for PaLM with different plan variants and sory prefixes. None is a shorthand for the prefix "Story:", Long abbreviates prefix "A native ⟨tgt_language⟩ speaker would write the story as:", Target is the translation of the prefix "Story:" in the target language, and Lang refers to "⟨tgt_language⟩ story:".

Entities	Story Completion
Una volta, c'era una mucca, un cane e una capra che andarono a visitare un altro villaggio. Decisero di prendere un autobus. Il cane era un po' preoccupato, ma la capra era molto felice e disse al cane: "Non preoccuparti! La gente sarà gentile con noi." Il cane non era sicuro, ma la capra era determinata. Così tutti e tre i loro amici salirono sull'autobus. Quando il conducente dell'autobus li vide salire, cominciò a ridere. "Siete voi tre amici che vi siete persi?" Disse il conducente . "Dove state andando?" La capra , che era più intelligente dei suoi amici , rispose: "Stiamo andando a visitare un altro villaggio." Il conducente dell'autobus si mise a ridere di nuovo. "Bene, amici! Non ho mai visto animali sull'autobus! Siete molto divertenti!" Disse. Il cane non era sicuro. La capra rispose: "Abbiamo pagato abbastanza soldi per prendere l'autobus." ⟨...⟩	Capra, Cane e Mucca erano grandi amici. Un giorno, presero un taxi per andare a fare un giro. Capra sedeva sul sedile posteriore, mentre Cane e Mucca sedevano sul sedile anteriore. Quando il taxi passava attraverso il villaggio, le persone dicevano: "Guardate! Guarda quel taxi! Che strano!" Capra, Cane e Mucca andarono a visitare un amico che viveva in una fattoria vicina. Quando il taxi arrivò alla fattoria, gli animali scesero e salutarono il loro amico. La fattoria era piena di erba fresca, così Cane, Capra e Mucca decisero di fermarsi un po'. Il taxi rimase fermo nella fattoria per sei ore. Intanto, i tre amici giocarono e giocarono e giocarono. ⟨...⟩

Plot Outline	3Act Structure
Una capra voleva andare a visitare un suo amico. Aveva un problema. Non aveva idea di dove fosse il suo amico. "Posso andare a visitare il mio amico?" Chiese alla mucca . "No," rispose la mucca. "Ma tu sai dove si trova." Disse la capra. "Sì," rispose la mucca. "Ma non ho tempo. Devo dormire." "Posso andare a visitare il mio amico?" Chiese alla capra al cane . "No," rispose il cane. "Ma tusai dove si trova." Disse la capra. "Sì," rispose il cane. "Ma non ho tempo. Devo giocare." ⟨...⟩	Una volta, la mucca, la capra e il cane viaggiarono insieme sul taxi. ⟨...⟩ Chiese la mucca. "Se non pagherai, il conducente del taxi chiamerà la polizia e sarai arrestata. Non puoi semplicemente scappare!" La capra non ascoltò e corse via. Il cane si precipitò verso il taxi per chiedere il resto all'autista. La mucca attraversò la strada con calma, perché aveva pagato il giusto prezzo. La capra uscì correndo dal taxi e scappò via. La capra non pagò il taxi. La capra fuggì. La mucca attraversò la strada con calma. Il cane tornò al taxi.

Table 15: Abridged examples of stories generated in Italian with different plans for story "Goat, Dog and Cow" (story ID: 4). Highlighted excerpts correspond to the events and/or characters mentioned in the input plan.