Dialogue & Discourse 10(1) (2019) 34–86 doi: 10.5087/dad.2019.103

A Narrative Sentence Planner and Structurer for Domain Independent, Parameterizable Storytelling

Stephanie M. Lukin[*] STEPHANIE.M.LUKIN.CIV@MAIL.MIL
U.S. Army Research Laboratory
Los Angeles, CA

Marilyn A. Walker MAWALKER@UCSC.EDU
Natural Language and Dialogue Systems Lab
University of California, Santa Cruz, CA

Editor: Vera Demberg

Submitted 07/2017; Accepted 04/2019; Published online 05/2019

Abstract

Storytelling is an integral part of daily life and a key part of how we share information and connect with others. The ability to use Natural Language Generation (NLG) to produce stories that are tailored and adapted to the individual reader could have large impact in many different applications. However, one reason that this has not become a reality to date is the NLG STORY GAP, a disconnect between the plan-type representations that story generation engines produce, and the linguistic representations needed by NLG engines. Here we describe Fabula Tales, a storytelling system supporting both story generation and NLG. With manual annotation of texts from existing stories using an intuitive user interface, Fabula Tales automatically extracts the underlying story representation and its accompanying syntactically grounded representation. Narratological and sentence planning parameters are applied to these structures to generate different versions of the story. We show how our storytelling system can alter the story at the sentence level, as well as the discourse level. We also show that our approach can be applied to different kinds of stories by testing our approach on both Aesop's Fables and first-person blogs posted on social media. The content and genre of such stories varies widely, supporting our claim that our approach is general and domain independent. We then conduct several user studies to evaluate the generated story variations and show that Fabula Tales' automatically produced variations are perceived as more immediate, interesting, and correct, and are preferred to a baseline generation system that does not use narrative parameters.

Keywords: Natural Language Generation, Personalized Storytelling, Sentence Planning

1. Introduction

Storytelling is an integral part of daily life and how we share information and connect with others. People often structure observed events into a story (Bruner, 1991; McAdams et al., 2006; Gerrig, 1993), so that an average day at work may later be described as a narrative where the events are exaggerated to revolve around the individual, rather than simply listing events that took place. In these natural story settings, stories may be told many times to different audiences but rarely told in the same way twice. A storyteller may explore different interpretations of the same incident from multiple points of view (Mateas, 2001), or use a richer style when telling a story to highly

∗. This work was done at the University of California, Santa Cruz.

Startled Squirrel	*The Fox and the Crow*
We keep a large stainless steel bowl of water outside on the back deck for Benjamin to drink out of when he's playing outside. The craziest squirrel just came by- he was literally jumping in fright at what I believe was his own reflection in the bowl. He was startled so much at one point that he leap in the air and fell off the deck. But not quite, I saw his one little paw hanging on! After a moment or two his paw slipped and he tumbled down a few feet. But oh, if you could have seen the look on his startled face and how he jumped back each time he caught his reflection in the bowl!	A Crow was sitting on a branch of a tree with a piece of cheese in her beak when a Fox observed her and set his wits to work to discover some way of getting the cheese. Coming and standing under the tree he looked up and said, "What a noble bird I see above me! Her beauty is without equal, the hue of her plumage exquisite. If only her voice is as sweet as her looks are fair, she ought without doubt to be Queen of the Birds." The Crow was hugely flattered by this, and just to show the Fox that she could sing she gave a loud caw. Down came the cheese,of course, and the Fox, snatching it up, said, "You have a voice, madam, I see: what you want is wits."

Table 1: *Startled Squirrel* and Aesop's *The Fox and the Crow*

interactive and responsive addressees (Thorne, 1987). When young adults describe situations in which their lives were threatened, they use different telling styles to convey different messages to the audience, such as empathy for others, preoccupation with one's own fear or sadness, or one's courage or bravery (Thorne and McLean, 2003). Retelling capabilities are showcased in Exercises in Style, where a sequence of simple events are told in 99 different ways (Queneau and Wright, 1981). Madden (2006) repeats the exercise in visual storytelling, creating different visual depictions of the same events in a story.

A computational treatment of storytelling should support the ability to retell stories in different ways, mimicking how human storytellers tailor their stories to the context and to their audience. Consider the personal narrative *Startled Squirrel* in Table 1 from the Spinn3r corpus of blogs (Burton et al., 2009). This telling is in the first person, using the narrator's own voice. The narrator tells about a time when they saw a curious squirrel in their backyard, who tried to drink out of a dog's water bowl. Upon getting closer to the bowl, the squirrel jumped at its own reflection and fell off the deck. This story primarily evokes a humorous response, however a different telling could instead evoke empathy for the squirrel if the story were told from the perspective of the squirrel itself, as depicted in the variation that our system can automatically produce, in Table 2. Similarly, Aesop's Fable *The Fox and the Crow* (Table 1), which is traditionally told from a third person perspective, could be framed from either the fox or the crow's perspective, affecting the reader's insight into each character's thoughts, as in the automatically produced variation in Table 2.

Startled Squirrel Variation Excerpt	*The Fox and the Crow* Variation Excerpt
I approached the bowl. I was startled because I saw my reflection. Because I was startled, I leaped. I fell over the deck's railing with my paw. My paw slipped off the deck's railing.	The crow sat on the branch of the tree. The cheese was in the beak of the crow. I observed the crow. I thought "I will obtain the cheese from the crow's beak!"

Table 2: Computational Variations for *Startled Squirrel* and Aesop's *The Fox and the Crow*

In order to computationally retell stories in different ways, the story representation must distinguish the content of the story from the telling. This distinction is classic in narratology and categorized as *fabula* and *sujet* (Propp, 1969). The *fabula* is comprised of the events in a story, represented abstractly as a set of building blocks that can be rearranged and from which more complex narrative forms can be built (Abbott, 2008). The *fabula* includes all the abstract components of the story world, including the characters, their goals, and the actions that take place in the story world.

On the other hand, the specific telling and framing of a subset of these events is the *sujet* (Propp, 1969). Constructing the *sujet* may include reordering or otherwise manipulating the presentation of events, choosing between a subjective or objective interpretation, the perspective from which the story is told, and the character voice, among other variations.

A computational storyteller requires a general representation of the *fabula* and a way to generate different *sujet* from a single *fabula*. One main limitation of work in this area to date, which has hampered progress in the field, is what is known as the Natural Language Generation (NLG) Story Gap, illustrated in Figure 1 (Lönneker, 2005; Callaway and Lester, 2002). The left-hand side of the figure is meant to depict the considerable line of research invested in the automatic production of different *fabula*, given a particular pool of content (Peinado and Gervás, 2006; Riedl and Young, 2010; Gervás et al., 2005) *inter alia*. This line of work has examined, for example, how plan-based approaches for selecting and ordering the content in different ways, e.g. selecting different events, leaving events out, or re-ordering events can have differential effects on the reader, such as enhancing the reader's feelings of suspense or surprise (Bae and Young, 2009; Ware and Young, 2011; Niehaus and Young, 2009). The NLG story gap refers to the fact that these story engines produce plan-like content representations, which unfortunately do not provide the information that is needed in order to render that content textually. This line of work often adopts a simple rendering strategy, defining templates by hand that directly realize each component of the *fabula*, as shown in the bottom left-hand-side of Figure 1. The result is that the only variations of the *sujet* that are possible in this approach are those that have to do with content selection, or those that are explicitly hand-crafted as template variations.

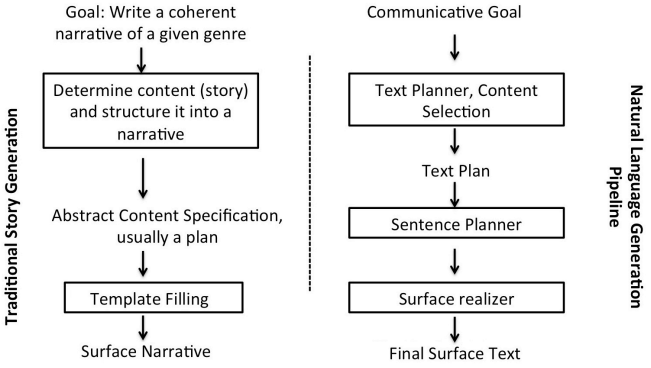


Figure 1: Differences between Story Generation and Natural Language Generation

In contrast, the right-hand side of Figure 1 shows the typical input to an NLG engine and the standard modular architecture that NLG uses to generate different textual renderings of their input. It should be possible, in principle, to generate different *sujet* by building on previous work in NLG, which provides an abundance of techniques for generating different textual variations from a meaning representation. However, NLG architectures assume that the input to the NLG is a text plan, whose leaves are syntactically and semantically grounded in linguistic representations. These linguistic representations are needed in order to apply sentence planning operations and produce the many different possible variations in texts for a fixed meaning representation, as we explain in more detail in Section 3. These representations are not compatible with story planning as they only contain information about a single sentence.

Thus the NLG story gap arises as the gap between the plan-like meaning representations used by story generators, and the input assumptions of NLG engines. Current practice is to fill this gap by hand, either by writing templates as shown in the left-hand side of Figure 1, or by constructing an NLG dictionary, which maps from each story meaning component to a linguistic form, such as a dependency tree, which NLG sentence planning operations such as aggregation and discourse structuring, can operate on (Mairesse and Walker, 2011; Callaway and Lester, 2002; Penning and

Theune, 2007). For example, Callaway and Lester (2002) mapped the story elements for *Little Red Riding Hood* into a syntactically grounded representation by hand, which then supported their work on the automatic generation of narrative variations. Similarly, work by Theune et al. (2007) on a system called *The Narrator* describes in detail how the story consists of causally related semantic story elements, and how each story element is mapped by hand to a dependency tree. Another approach to bridging the NLG story gap recently proposed by Concepción et al. (2016b) suggests that the use of a controlled language (Controlled Natural Language) for specifying story content could make it simpler to map story plan structures to syntactic patterns in a very general way. These syntactic patterns could in principle then be converted into the linguistic representations needed by different NLG engines (Concepción et al., 2016a,b,c).

In sum, to date, bridging the NLG story gap has required a considerable amount of hand-crafting for **each story** that a storytelling system would want to tell, and this has prevented narrative systems from generating rich and diverse variations over a variety of different topics and genres. Our approach to this problem consists of three separate contributions that together form the Fabula Tales storytelling system:

1. We propose a particular take on bridging the NLG story gap that takes existing stories from different genres, creates a *fabula* representation using *Scheherazade*, an easy-to-use annotation tool, and then automatically maps the *fabula* to a general NLG representation;

2. We develop an NLG engine with narratologically inspired sentence planning parameters and show how we can generate different tellings of stories using a narratological structurer that employs these parameters for any existing story that has been annotated;

3. We evaluate different generated tellings at both the sentence and the story level for different evaluation criteria.

Bridging the NLG Story Gap. In order to bridge the NLG story gap, we first require the preservation of content representing the *fabula* or semantics of a narrative, and the creation of linguistic representations that can be used to generate tellings, *sujet* or syntactics (Section 3).

We use existing stories, in their textual form, that come from different genres and different topics, as exemplified in Table 1. Our story corpus selection allows us to explore the domain-independence of our approach: the corpora consists of 36 Aesop's Fables and 108 first person social-media blogs from the Spinn3r corpus (Burton et al., 2009), two radically different genres. We adopt Elson's Story Intention Graph (SIG) (Elson, 2012a), as the representation of *fabula*. In addition to its strong theoretical motivation, one advantage of using the SIG as a *fabula* representation is that it can be produced in a lightweight way (only one to two hours per story) using a corresponding annotation tool called *Scheherazade* that supports the production of *fabula* by annotation of texts (Elson and McKeown, 2009).

We then create a general model that maps from the SIG to novel Lexical-Semantic Story Trees (LSSTREES) containing linguistic representations needed for NLG, building on our prior work (Rishes et al., 2013). This highlights a second advantage of the SIG representation: annotation using *Scheherazade* maps each predicate and constant in the story's logical representation to a lexical item from the off-the-shelf lexical resources, VerbNet (Kipper et al., 2006) and WordNet (Fellbaum, 2010). This grounding to lexical items (with their subcategorization frames) allows us to create the general mapping model compatible with all story domains. We show that the mapping produces good quality Lexical-Semantic Story Trees and generates good baseline stories, without

expert handcrafting.

Generating Stories with Narrative Variations. After showing that we can develop a general representation of *fabula* for any story domain, our second contribution is to generate different *sujet* by implementing narratological sentence planning and a set of discourse relations (Section 4), and a story-level narratological structurer that sits on top of the Lexical-Semantic Story Trees (Section 5).

The Lexical-Semantic Story Trees are manipulated by Fabula Tales' narrative sentence planner, based on the architecture of the PERSONAGE expressive Natural Language Generation engine (Mairesse and Walker, 2008, 2011), with parameters inspired by theories of narratology (Genette and Lewin, 1983; Prince, 1974; Lönneker, 2005; Bal, 1997). Our narrative sentence planner supports changing point of view (first or third), inserting direct speech acts, and supplementing character voice using operations for lexical selection, discourse structuring, and pragmatic marker insertion.

We develop a narratological planner for Fabula Tales that operates above the narrative sentence planner. Our narratological planner is not a narrative content planner; thus our approach assumes that all content from the Story Tree will be told, and the planner determines which narrative parameters should be applied to generate the story. The narratological structurer determines variations at the story-level by focusing on the entire flow of the story, rather than just at the sentence level. Training data is obtained by overgenerating different variations of sentences on a sentence-by-sentence basis. The sentences are then ranked by subjects using a novel Create-Your-Own-Story annotation paradigm, to learn the impact of each narrative parameter.

Evaluation of Narrative Theories on Generated Stories. By combining these parameters, the generated variations evoke diverse framing and voice alterations at both the sentence and story level. We explore how different narrative parameters lead to different perceptions of the story, evaluating on holistic narrative metrics of immediacy, interest, correctness, and preference. We use this experimental data as input for a classification experiment where we rank possible choices that the generator can make in terms of which sentences are the most impactful. We conclude in Section 6 where we discuss limitations, future work, and applications.

2. Related Work

2.1 Narrative Variation in Storytelling Systems

The storyteller has many devices at their disposal to frame stories, including changing the overall tone, mood and effect of the story to distinguish between "who sees?" and "who speaks?" (Genette and Lewin, 1983). Theories of narratology provide a number of narrative devices or parameters to produce diverse framings of a *fabula* (Bal, 1997; Lönneker, 2005; Genette and Lewin, 1983; Prince, 1974). Lönneker (2005) categorizes parameters into three broad categories of *Time*, *Mood*, and *Voice*, each with sub-categories, as detailed in Table 3. Narrative variations to *Time* and *Mood* primarily involve forms of narrative content planning, that is, determining or generating the events and the structure in which to tell (*fabula*). The *Voice* parameters influence the realization (*sujet*).

We study the NLG story gap and a systems' ability to generate these diverse tellings. We expand the gap presented in Figure 1, positing that a storytelling architecture has the potential for four gaps to arise, as we depict in Figure 2. The first gap occurs when story content for the storytelling pipeline is difficult or time consuming to create. Closely linked to the first gap, the second gap

Parameter	Explanation
Time: Order	Sequence in which events are told, in comparison with the sequence in which they "actually happened". In synchrony, the event sequence in discourse corresponds to the sequence of the story. Anachronies can take the form of flashbacks (retrospectives) or flashforwards (anticipations).
Time: Speed	Relation between story time and discourse time. Congruence exists probably only in single scenes; otherwise timelapses (accelerations), time jumps (ellipsis), time expansions (decelerations), or pauses are used to achieve different degrees of explicitness and emphasis.
Time: Frequency	Relation between the number of times a (similar) event happened, and the number of times an event is told. The following realizations are distinguished: singulative (one-to-one relation), repetitive ("recount several times what happened once"), and iterative ("recount once what happened several times").
Mood: Distance	Combination of amount of information conveyed and narrator intrusion. Stereotypically, detailed information and low narrator participation indicate imitation or "direct" dramatic mode, as opposed to a "distant", mediated narrative mode. This parameter also affects the way in which speech is reproduced.
Mood: Focalization	Accessibility of knowledge needed to select story events for presentation in discourse. If a narrative instance disposes of unrestricted knowledge of the story world, it uses external focalization; if the knowledge is restricted to a character's field of perception, focalization is internal.
Mood: Point of View	Spatial, temporal, and ideological points of view from which events are described. Events can be described from the point of view of different characters. This parameter covers more aspects than focalization.
Voice: Time	Time relation of the narrating action to the story event. Events can be told while they are happening (concurrently), retrospectively, or prospectively.
Voice: Person	Narrator participation. A homodiegetic narrative instance is a character of the current narration (grammatical realization typically in the first person), while a heterodiegetic narrative instance is "absent" from the current narrative and not referred to. In a second-person narrative, the protagonist is the reader.

Table 3: Narratological Parameters presented in Lönneker (2005)

occurs during the process of applying a transformation to the content in order to make it compatible with the storytelling system. Below, we discuss story planners that require dependency information between plot points but lack accessible tools for their creation. We also review NLG engines that require detailed syntactic structures as input. The result is that new stories from different genres are more difficult to create. The work we present in this article does not face these two gaps because the intermediary story representation we use, the SIG, is easily obtainable from any genre of natural, unstructured story texts using its accompanying creation tool, *Scheherazade*. In Section 3, we describe how this intuitive tool has a low authorial burden for creating new content.

The third gap occurs when neither the content pool (the selected *fabula*) nor its representation are rich enough for story planning. This tends to occur in systems that make use of syntactic representations which are ripe for narrative manipulation, but do not necessarily preserve story-level information, making it difficult to alter the story when nothing is known beyond a single sentence, as we discuss below. In our work, we develop a novel representation, Lexical-Syntactic Story Trees (LSSTREES), that preserve story-level discourse information across story predicates (Section 3).

The LSSTREES are compatible with content planning as well as syntactic transformations in sentence planning.

Finally, the fourth gap precludes variation of the *sujet* due to a lack of syntactic information (rather than semantic information in the third gap); a system is therefore unable to dynamically alter the telling once a story point has been selected, instead making use of templates or fixed-strings. The remainder of this section discusses related work in computational storytelling within the scope of the narrative parameters described above and limitations to the work with respect to the four elements of the NLG story gap presented here.

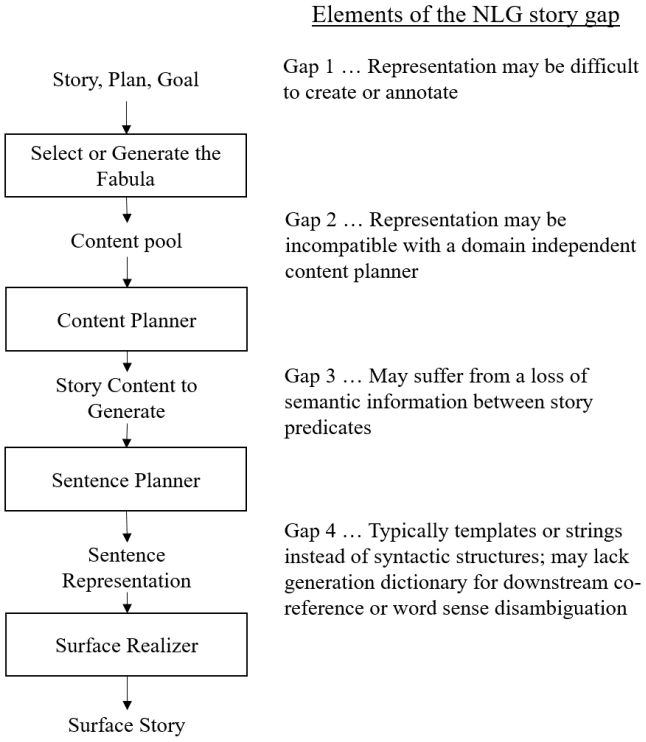


Figure 2: Gaps between story representation, story generation, and NLG

As stated above, Fabula Tales does not perform any temporal content planning (i.e., *Time:Order, Time:Speed, Time:Frequency*) and assumes every story point will be selected and told in the order in which it occurred. There are many approaches to story content selection, a number of which select content to prioritize author-level goals. In these systems, broad narrative goals are defined by hand prior to story generation, and then the assertions of the story world are reasoned over to assign the *Time* of the narrative, using, for instance, hierarchical task networks (Lebowitz, 1983), case-based reasoners (Peinado and Gervás, 2006; Turner, 1993; Gervás et al., 2005), or bipartite or tripartite optimizations (Winer and Young, 2016; Barot et al., 2015). Other systems take a character-driven approach to story content planning and select content based on character knowledge, goals, and their relationships with others in the story world (Meehan, 1977; Riedl and Young, 2010; Theune et al., 2004). Stories are generated when the preconditions of author or character goals are met. Despite control over the story generation, these systems place less priority on the realization of the texts, and are instead typically use pre-authored pieces of story text or templates.

Mood:Focalization is a narrative device that has been explored using planning engines that identify which events or inner thoughts characters are aware of at the moment, including character's beliefs, desires, hidden intent, or inner states of mind. These affect the selection of the events the planner selects to tell; for example, stories with surprise endings are generated by exploiting the disparity of knowledge between a story's reader and its characters (Bae and Young, 2009; Bae et al., 2011) while others create conflict (Ware and Young, 2011, 2012; Ware et al., 2014; Niehaus and Young, 2009). Curveship enables similar flexibility of story framing of focalizations or temporal orders, or the speed of the narrative (*Time:Speed*) (Montfort, 2007, 2009), although Curveship is

only applicable to the Interactive Fiction domain, and similar to the work described above, only offers a templatic realization strategy.

On the opposite side of the NLG story gap are prose generation systems that prioritize diverse *sujet* generation, typically effecting the *Mood:Distance* and *Voice:Person* of the resulting narratives. These works use rich and flexible syntactic representations that can be manipulated. Some approaches include the use of document plans and dependency trees (Theune et al., 2007), or annotated source material (i.e., story texts) as Controlled Natural Language, which restricts grammar and vocabulary to afford flexibility for sentence planning (Concepción et al., 2016a,b,c), or a narrative plan compatible with the off-the-shelf generator FUF-SURGE (Callaway and Lester, 2002; Elhadad and Robin, 1996). Shortcomings of these carefully crafted approaches tend to be the limited domain representations and the time constraint to manually create these representations, e.g., Callaway and Lester (2002) can only generate stories in the Little Red Riding Hood domain, and both approaches require the manual construction of a formalism representing the characters, story assertions, and various parameters, whereas our approach offers an intuitive user interface for defining these story requirements.

Recent work performs both story planning and diverse text generation using data-driven approaches that make use of large corpora of unstructured texts, rather than carefully curated content or syntactic representations as input. One interactive approach takes turns with the user to co-construct a story using data from the Spinn3r blog corpus or from movie scripts (Swanson and Gordon, 2008; Munishkina et al., 2013). After the user types the next sentence of a story, the algorithms search for similar sentences from stories in the corpus using Term Frequency-Inverse Document Frequency (TF-IDF) or other search criteria. Returning the next sentence from the selected story to the user progresses the co-constructed narrative forward. The Scheherazade story generation system (not to be confused with the *Scheherazade* annotation tool for SIGs), learns causal graphs from texts obtained by crowd-sourced workers prompted to write a short story about a particular topic (Li et al., 2013; Li, 2015). The planner performs well because the texts are constructed from a prompt and assume a causal and event-centric structure. It remains to be seen how this approach would apply to narratives collected "in the wild" such as in the Spinn3r corpus, a large portion of which contain orientation and evaluation segments prevalent in oral narratives (Labov and Waletzky, 1997; Rahimtoroghi et al., 2013, 2014). These textual learning approaches are advantageous for domain independence and can retrieve different styles of prose by mining a corpus. However, because these algorithms are retrieval-based, when they find a matching response in the corpus, the algorithm returns the response as-is without varying the selected text. Therefore, these systems' expressivity with respect to language generation are restricted to the original narrative text or script.

Other joint story planning and realization systems utilize Recurrent Neural Networks or Long-Short Term Memory Convolutional Sequence-to-Sequence neural networks (Roemmele, 2018; Fan et al., 2018; Peng et al., 2018). These approaches similarly enable the learning of many topics and genres and have the additional advantage of generating text learned from the corpora. Yet to date, these models do not attempt to diversify the narrative texts according to any theories of narrative, but rather posit that variations can be learned inherently from the datasets. These approaches construct a text word-by-word, sentence-by-sentence, but these texts do not model the overall narrative scope of the *fabula*, or offer diverse sentence planning for generating different *sujet*. To date, these neural-based approaches do not afford these storytelling aspects that more traditional approaches have done in the past.

Tasks for evaluating the consistency of stories show promise for someday being used to evaluate causality in automatically generated stories. Hu et al. (2013) and Rahimtoroghi et al. (2016) learn event pairs from the Spinn3r corpus using unsupervised approaches and causal potential. Another evaluation, the Corpus of Plausible Alternatives (COPA), is created from hand-annotated causality pairs from Spinn3r, and the task is to select the most likely event to occur next in the story (Roemmele et al., 2011). Mostafazadeh et al. (2017) creates a synthetic dataset to measure the predictability of subsequent events, but unfortunately, a bias was identified in the dataset creation; as a result, simple natural language processing tricks can obtain a high score on the task, rather than examining the content itself (Srinivasan et al., 2018; Sharma et al., 2018).

A new task of visual storytelling has been driven forward by improvements to computer vision algorithms. This form of storytelling faces the same challenges as text-based story generation, but an additional challenge is that the *fabula* must be determined from computer vision algorithms. Huang et al. (2016) create a corpus for visual storytelling (VIST), yet the guidelines for the collection effort do not explicitly take into consideration the entire story as a whole, nor do they explore diverse narrative variations; the stories human annotators create about a sequence of images tend to be shallow and action-oriented. The Visual Storytelling Challenge[1] is the first shared task in visual storytelling, and they, as well as Wang et al. (2018), offer several subjective evaluation metrics that are applicable for text-based storytelling as well (e.g., "focus" and "expressiveness"). Lukin et al. (2018) poses challenging questions for the visual storytelling task in order to bridge this modern task to its roots in narrative theories, including specifying that visual story generation systems must be flexible in both content planning and realization of different narrative goals and be able to adapt the narrative to the audience.

2.2 Foundations of the Fabula Tales Storyteller

Fabula Tales' automatic syntactic translation from SIG to LSSTREE builds upon previous work first described in Rishes et al. (2013). However, that work only explored a single domain, Aesop's Fables, whereas the model presented in this article has been improved and tested on an additional 108 stories from the personal blog domain. New evaluations are presented that measure the quality of the baseline translation algorithm in terms of text similarity, semantic text similarity, fluency, and grammaticality. Furthermore, the system presented here models semantic and discourse relations between plot elements, which the original method did not model. This article reviews the syntactic translation process alongside the new contributions of this article in order to present together the complete storytelling pipeline.

Fabula Tales' narrative sentence planner implements similar parameters and linguistic representations as the PERSONAGE expressive NLG engine (Mairesse and Walker, 2008, 2011). PERSONAGE manipulates Deep Syntactic Structures (DSYNTS) (Lavoie and Rambow, 1997) according to parameterized models grounded in the Big Five personality traits, providing a large range of pragmatic and stylistic variations of a single utterance. In PERSONAGE, the style to be conveyed is controlled by a model that specifies values for different stylistic parameters (such as verbosity, syntactic complexity, and lexical choice). PERSONAGE requires hand crafted text plans and DSYNTS, limiting not only the expressiveness of the generations, but also the domain. PERSONAGE has been used as a way to help authors to reduce the authorial burden of writing dialogue instead of relying on scriptwriters for games (Reed et al., 2011), but still relies on hand-authoring DSYNTS. Fabula Tales introduces

1. http://www.visionandlanguage.net/workshop2018/#challenge

the first tool for automatically creating DSYNTS, which allows for PERSONAGE's sentence planning to be repurposed for the narrative space.

Another source of linguistic variation supported by PERSONAGE and reimplemented by Fabula Tales is splitting and aggregating sentences at the discourse level. Aggregation operations help to avoid repetition and produce more coherent, concise, and context aware output (Cahill et al., 2001; Scott and de Souza, 1990; Paris and Scott, 1994). Several NLG systems use Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) for aggregation and sentence planning (Walker et al., 2007; Howcroft et al., 2013). Aggregation is inclusive of aspects of abstractive text summarization, with parallels in sentence compression, fusion, lexical paraphrasing, and reorganization by reducing syntactic structures through the removal of articles (Grefenstette, 1998), manipulation over syntactic trees (Knight and Marcu, 2000), dependency trees (Filippova and Strube, 2008), or grammar (Riezler et al., 2003) to name a few. Not all text summarization operates over syntactic units, instead employing text-to-text generation (Chandrasekar and Srinivas, 1997; Knight and Marcu, 2002; Marsi and Krahmer, 2005). However, recent work introduces a task based on starting with small meaning-representations and recombining them in different ways (Narayan et al., 2017), a step towards joining the advances and contributions of aggregation and text summarization methodologies.

Our method for training our narratological structurer is based on previous work on overgenerate and rank. NLG systems have made use of the overgenerate and rank methodology in which a variety of sentence variations are generated from a model that first overgenerates, and then ranks the generated output based on some measure of "goodness" relevant to the task. Previous work has used statistical models as an objective scoring function to measure a set of generated candidate utterances based on a variety of features. Simple ranking based on n-grams are used to generate from Abstract Meaning Representations (Langkilde and Knight, 1998; Langkilde-Geary, 2002) and to generate dialogues with alignment and personality cues (Isard et al., 2006). Lexical and conceptual similarity scores influence near-synonym selection (Inkpen and Hirst, 2004) and correctness and grammaticality scores influence sentence generation (Gardent and Kruszewski, 2012). The overgenerate and rank methodology has been shown to be useful for error mining with human judgment as a scoring metric and for future parameter adjustments and feedback (Walker et al., 2002, 2007; Mairesse and Walker, 2010b; Walker et al., 2013; Gardent and Kruszewski, 2012). In the narrative space, overgenerate and rank has been used to combine a rule-based overgeneration phase with a statistical ranking phase by probabilistic parsing to rank sentences in a story for its naturalness (Ahn et al., 2016), yet this work does not employ any narrative specific sentence planning; the highest ranked variant for each sentence in the story is simply concatenated to the other selected sentences.

3. Bridging the NLG Story Gap

This section describes how Fabula Tales bridges the NLG story gap. As *fabula*, we use existing stories about different topics and from different genres. The strength of our approach is that it allows us to test whether our methods for generating different *sujet* can be applied across many different *fabula*. A limitation of our approach is that we only have a fixed set of story points within each *fabula*, in contrast to work on story generation systems whose focus has been to explore story variations that result from manipulation and selection of events from the *fabula* (Bae et al., 2011; Riedl and Young, 2004; Gervás et al., 2006).

Fabula Tales' bridge is depicted in Figure 3. It begins with the raw text from an existing story, and produces a representation upon which the narrative sentence planner and narratologi-

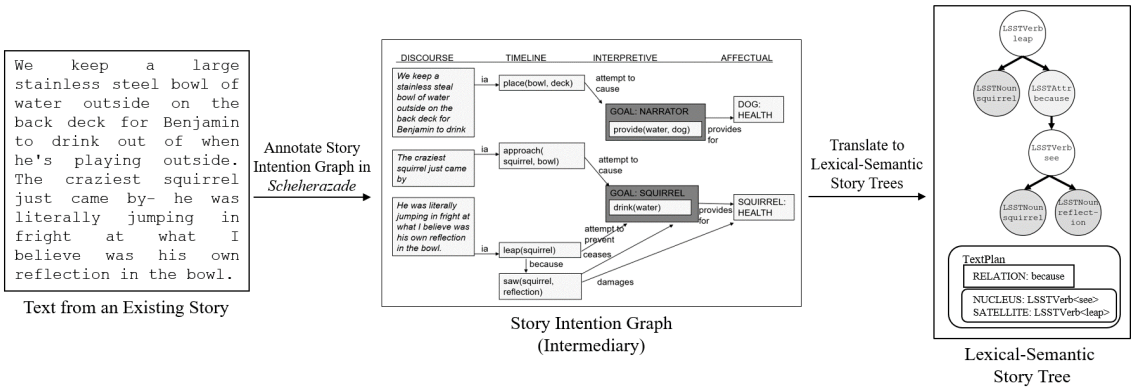


Figure 3: Fabula Tales translation from text, to intermediary Story Intention Graph, to Lexical-Semantic Story Trees

cal structurer operate. As the figure shows, the story content is first manually annotated using the *Scheherazade* annotation tool, in order to produce an intermediate representation of the *fabula* as a Story Intention Graph (SIG) (Elson, 2012a). The SIG represents a story along a variety of dimensions, including plans, goals, and actions of characters. This formalism emphasizes the key elements of a narrative rather than attempting to model the entire semantic world of the story. The SIG representation is then automatically translated to Lexical-Semantic Story Trees (LSSTREES). LSSTREES consist of a syntactic and semantic component. The linguistic representation can be directly converted to Deep Syntactic Structures (DSYNTS) in order to vary sentences at the syntactic level using the off-the-shelf surface realizer REALPRO (Lavoie and Rambow, 1997; Mel'čuk, 1988). The semantics are captured through text plans, and model three discourse relations: CONTINGENCY, TEMPORAL ORDER, and ATTRIBUTION.

Section 3.1 describes the SIG formalism. The creation of SIGs requires **no** parsing or preprocessing of the input; the story text is manually annotated using a tool called *Scheherazade* (Elson and McKeown, 2009). *Scheherzade* is a user-friendly annotation tool that is generalizable and does not require specific domain or knowledge or deep linguistic knowledge or a generation dictionary. Word sense disambiguation is done as part of the annotation process: all of the predicate-argument structures representing the structure of the story and the character intentions are lexically grounded in either WordNet synsets (Fellbaum, 2010) or VerbNet verb nodes (Kipper et al., 2006). The DramaBank, an existing corpus of Aesop's Fables annotated as SIGs (Elson, 2012b), was first used to explore our translation pipeline. We then test the domain and genre independence of the pipeline by using a second corpus of first person informal blogs, PersonaBank, that we created in prior work (Lukin et al., 2016). Because it is used extensively here in our experiments, we describe PersonaBank and summarize the SIG annotation process. Previous work, as well as our own, has shown that the *Scheherazade* annotation tool can be used by non-expert annotators and does not require a background in linguistics or computer science, nor are annotators required to be domain experts, whereas other work, as we described in the previous section, may require careful hand-authoring of syntactic structures or detailed domain knowledge.

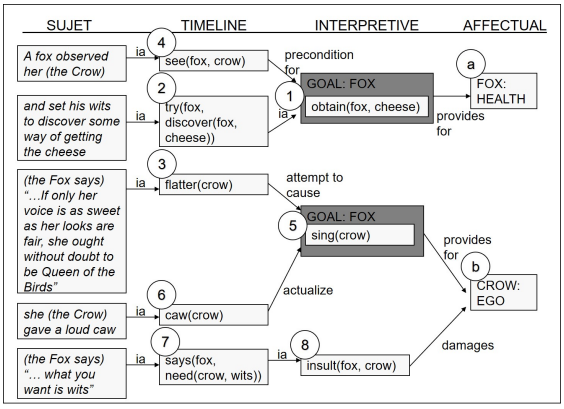
The story generation process is streamlined after a SIG is created. Discourse relations annotated in the SIG are used to construct text plans in the LSSTREES, and a syntactic representation of each story point is automatically translated from the SIG. Section 3.2 shows how this semantic

and syntactic translation process is fully automated and does not require construction of semantic or syntactic structures by hand or by using templates. This final representation allows the narrative sentence planner and narratological structurer to generate different narrative variations of a story. We show how following the pipeline in Figure 3 generates a baseline text without narrative variations, which we use in subsequent evaluations, and present evaluations verifying the fidelity of the translation process using this baseline text (Section 3.4).

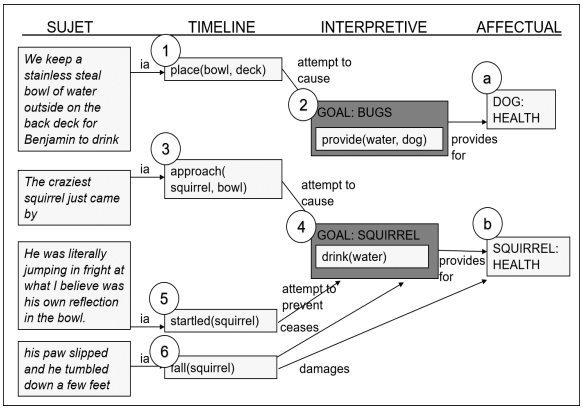
3.1 Story Intention Graphs: Intermediary Fabula

The SIG formalism is a computational model of narrative that goes beyond the surface form of a text, as opposed to style. SIGs separate "what the story *is* [*fabula*]" from "how the story is *told* [*sujet*]" (Elson, 2012a). These encode a single *sujet* from which the underlying *fabula* is derived. In contrast, the primary goal of this work is to transform a single *fabula* into many *sujet*. As is the nature of the *sujet*, a telling is only one interpretation or rendering of a larger narrative discourse. For a particular SIG, some events may not have been made explicit in the original *sujet*, and thus are excluded from the derived *fabula*. However, the assumptions and inferences that a reader makes to interpret the story can be added to the SIG in one of its deeper semantic layers, as we explain below.

The first step in annotating a story as a SIG is to define all of the characters and props. These are given unique IDs, and are defined by identifying a WordNet synset that is of the right type for that character or prop, e.g. *a fox* or *a tree*. Then, the events of the story timeline are defined and their propositional representations use these character and prop entities as arguments. A story in the SIG formalism is represented by four layers as in the example SIG for *The Fox and the Crow*, shown in Figure 4a: the *sujet* or TEXTUAL LAYER, the TIMELINE, the INTERPRETIVE layer, and the AFFECTUAL layer. The nodes in each layer are connected by arcs signifying semantic or discourse relationships between the nodes, within or across layers. The original story, the *sujet* (first column in Figure 4a) is first divided by the annotator into textual segments, where each segment, in the annotator's view, represents a distinct, coherent story point.



(a) A Story Intention Graph for *The Fox and the Crow*



(b) A Story Intention Graph for *Startled Squirrel*

The other three layers of the SIG comprise different elements of the *fabula*, each layers' nodes derived from VerbNet frames with WordNet story elements. The WordNet and VerbNet senses are utilized in the LSSTREE creation process to build a generation dictionary for downstream word sense disambiguation and co-reference resolution upon text realization. The TIMELINE layer summarizes the actions and events that occur. The INTERPRETATION layer captures story meaning

derived from agent-specific plans, goals, attempts, outcomes and affectual impacts, and the annotators' interpretation of *why* characters were motivated to take the actions they did, adopting a "theory of mind" approach to modeling narratives (Palmer, 2007). The final dimension is the AFFECTUAL layer, representing deeper motivations underlying character goals and the effect these goals have on the characters. There are 12 basic types of affect, including *health*, *ego*, *wealth* as described in more detail in Elson and McKeown (2009).

The fact that the *Scheherazade* annotation tool can be used by non-expert annotators to easily create new SIG story encodings was first demonstrated by Elson's work on the creation of the DramaBank corpus of SIGs (Elson, 2012b). Our work uses a subset of DramaBank consisting of all 36 Aesop's Fables, such as the example *The Fox and The Crow* shown in Table 1, and other well-known stories like *The Boy who Cried Wolf* and *The Fox and the Grapes*. A simplified SIG for *The Fox and the Crow* is shown in Figure 4a. Numbers indicate TIMELINE or INTERPRETATION events, and letters label the AFFECT nodes. The SIG specifies that the Fox's goal (#1) is to obtain the cheese from the crow. This would *provide for* his health, (A), represented as an AFFECT node. When the fox *sets his wits to discover some way of getting the cheese* this is encoded by the annotator as *the fox tries to discover how to obtain the cheese* (#2) which is *interpreted as* (ia) his goal (#1). A *precondition* arc is created to restrict that the goal of obtaining the cheese can only be initialized if the fox has first seen the crow (#4). The fox also has a goal (#5) that the crow will sing. By flattering the crow (#3) the fox *attempts to cause* (achieve) that the crow will sing. If the crow caws (#6), this would *actualize* the goal of the crow singing. The singing itself *provides for* the Crow's ego, (B), represented as an AFFECT node. When the fox says *what you want is wits* this is encoded by the annotator as *the fox said the crow needed wits* (#7) which is interpreted as the fox insulting the crow (#8). This *damages* the Crow's ego (B).

Our construction of the PersonaBank corpus is a second demonstration that *Scheherazade* can be used to create SIGs for stories with a variety of author styles and topics. PersonaBank is a corpus of 108 SIGs for blog stories from the Spinn3r corpus, which includes the *Startled Squirrel* story from Table 1 (Lukin et al., 2016). The SIG for the *Startled Squirrel* is shown in Figure 4b. Again, numbers indicate TIMELINE or INTERPRETATION events, and letters label the AFFECT nodes. The narrator places a bowl on the deck (#1) as an *attempt to cause* the goal of the narrator to give the dog some water (#2) which would *provide for* the dogs' health (a). Then the squirrel approaches the bowl (#3) as an *attempt to cause* (achieve) the squirrel's goal to drink the water (#4) which would *provide for* the squirrel's health (b). When the squirrel is startled (#5), this *attempts to prevent* (blocks) the goal of drinking the water, and when the squirrel falls (#6) this both *ceases* the goal (#4) and *damages* the squirrel's health (b).

Scheherazade provides a built-in generation module as part of the annotation process so that the annotator can see a realization of the underlying representation in real time as they annotate in order to verify that the underlying representation being constructed is what the annotator intends (Bouayad-Agha et al., 1998; Elson and McKeown, 2009). We will call this the *Scheherazade* realization. Table 4 shows the original story and the *Scheherazade* realization for the *Startled Squirrel*. *Scheherazade* uses templates and lexical realizations from the WordNet nouns and VerbNet frames to directly realize the underlying SIG semantics, without attempting to produce any type of variation in its realizations. Instead, it produces text in a fixed way for the selected encoding, e.g., the SIG semantics "APPROACH(SQUIRREL, BOWL), CAUTIOUSLY" will always be realized as *The squirrel cautiously approached the bowl.* Furthermore, there are odd and redundant phrasings in the *Scheherazade* realization because of its templates, for example, *The second squirrel leaped*

Startled Squirrel	*Scheherazade* realization
We keep a large stainless steel bowl of water outside on the back deck for Benjamin to drink out of when he's playing outside. The craziest squirrel just came by- he was literally jumping in fright at what I believe was his own reflection in the bowl. He was startled so much at one point that he leap in the air and fell off the deck. But not quite, I saw his one little paw hanging on! After a moment or two his paw slipped and he tumbled down a few feet. But oh, if you could have seen the look on his startled face and how he jumped back each time he caught his reflection in the bowl!	A narrator placed a steely and large bowl on a back deck in order for a dog to drink the water of the bowl. A squirrel approached the bowl. The squirrel began to be startled because it saw the reflection of the squirrel. The squirrel leaped because it was startled and fell over the railing of the deck and because it leaped. The squirrel held the railing of the deck with a paw of the squirrel. The squirrel fell, and the paw of the squirrel slipped off the railing of the deck.

Table 4: *Startled Squirrel* and a *Scheherazade* realization

Topic	Excerpt from Original Story	*Scheherazade* Realization
Wildlife, Bugs	Lillian found a wasp on the window at the farm.	A girl named Lillian found a wasp on a window of a farm.
Holidays, Christmas, Family	We tied [the christmas tree] down to the roof and go get hot chocolate.	The group of relatives of the narrator tied the pine tree onto the roof of the car. The group of relatives of the narrator drank some cocoa.
Romance, New Romance	I took a few pics and was blown away by the beauty of my girls.	The narrator photographed the bride and the maid of honor and noticed that the bride and the maid of honor was gorgeous.
Family	The trip started with a much anticipated but never duplicated dinner at Rainforest Cafe.	The family of a narrator ate at a restaurant named Rainforest Cafe.
Everyday Events, Technology	So he wanted to get another [phone]	The father of the narrator wanted to acquire a new second (#2) telephone.
Pets, Everyday Events	We went to a no-kill shelter to get our first cat.	The husband of the narrator and the narrator went back to a humane shelter in order to adopt a cat.

Table 5: Excerpts from PersonaBank, illustrating *Scheherazade* realizations

because it was startled and fell over the railing of the deck and because it leaped. Because the *Scheherazade* generator focuses on semantic fidelity to the SIG, we use it below as a baseline for measuring whether our translator bridge preserves story content (Section 3.4).

The primary motivation behind the creation of PersonaBank was to test the use of the SIG as a representation of *fabula* in our pipeline. In general, we selected stories that had a clear sequential timeline, and most of the selected stories are shorter than 300 words, with the minimum and maximum number of words to be 104 and 959 respectively (Table 6). Trained annotators[2] can annotate the timeline layer of a story in about one hour. Annotating the interpretive and affectual layers requires more subjective judgment and takes an additional hour for each story. As shown in Table 6, all stories were annotated with the timeline layer, 21 of which were annotated with the interpretive layers. Each story was annotated by a single annotator, thus the SIGs represent one interpretation of the story, one possible *fabula*.

2. Annotators of PersonaBank were undergraduate research assistants associated with the Natural Language and Dialogue Systems Lab at the University of California, Santa Cruz.

Sample excerpts from PersonaBank stories illustrating the range of topics covered and their *Scheherazade* realizations are shown in Table 5. Several of these story excerpts illustrate how the first-person *I* is typically mapped to a character called *the narrator* in the SIG encoding: it is not possible to use deictics like *I* or *me*, or anaphors like *it, he, she*. Because the SIG representation uses unique IDs for each character in a story, our realization engine can easily replace definite references like *the narrator* with other ways of referring to the same character. For more details of the corpus and its creation, please see Lukin et al. (2016).

Statistics	Stories
Total stories	108
Positive stories	55
Negative stories	53
INTERPRETATION layers annotated	21
Avg. length (# words)	269

Table 6: PersonaBank Statistics

3.2 Lexical-Semantic Story Tree Translation

We define Lexical-Semantic Story Trees (LSSTREES) as structures that contain both semantic and lexical information about a particular story point from a SIG's TIMELINE layer, including the action and actors involved in a particular story point, syntactic representation of these semantics, and text plans that detail the discourse relations that hold between propositions within the story. These structures are operated on and manipulated by the sentence planner (Section 4) and narratological structurer (Section 5) in order to produce narrative variations. Natural language text can be directly realized from these LSSTREES by a surface realizer, thus these structures provide the storytelling system with both semantic knowledge needed for the story *fabula* and syntactic information about the content to generate different *sujet*.

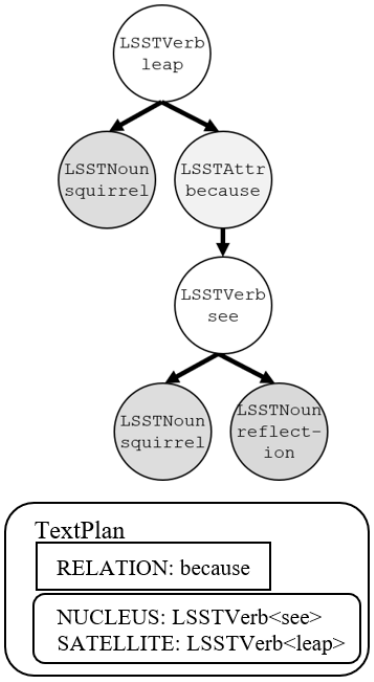


Figure 5: Lexical-Semantic Story Tree

Information from each SIG story point is automatically extracted and organized into LSSTREES, depicted in Figure 5. This translation does not use the original story text, but only the SIG structures; no parsing of text is required. Instead, we develop a mapping of syntactic structures that corresponds roughly to parts of speech: verbs, nouns, adjectives, and prepositional phrases. These parts of speech are defined in the SIG because part of the annotation process involves lexically grounding each predicate and constant for story events in either WordNet or VerbNet. These lexical resources provide both part of speech information as well as sub-categorization frames. Word senses are preserved in the mapping to LSSTREES in order to support word sense disambiguation and lexical choice in sentence planning. Finally, each story entity (characters or props) has a unique identifier, so there is no need to perform co-reference resolution. These structures and this syntactic translation was first developed in Rishes et al. (2013).

Content Assertions
1: assert(approach (*squirrel*, *bowl*))
2: assert(saw (*squirrel*, *reflection*))
3: assert(leap (*squirrel*))
4: assert(fell (*squirrel*))

Figure 6: Content assertions from the *Startled Squirrel*

Text plans that preserve a key set of discourse relations are constructed from the SIG, based on the Penn Discourse TreeBank (PDTB) discourse relations (Prasad et al., 2008). The annotation of a story requires story events to be ordered on the SIG timeline: these relations are represented as the PDTB TEMPORAL ORDER relation. We also represent the ATTRIBUTION relation between a speaker and their utterance in order to later vary whether an utterance is realized as direct or indirect speech. Finally, because causality is considered a key relation in the structuring of narratives, we represent the PDTB CONTINGENCY relation, an extremely common relation in PersonaBank.

A subset of content assertions from the *Startled Squirrel* are seen in Figure 6. A CONTINGENCY relation between assertion 1 and assertion 2 could result in the textual realization of *The squirrel saw its reflection because it approached the bowl*, and the TEMPORAL ORDER relationship between assertions 3 and 4 could yield *The squirrel leapt. He fell.* The derivation of the discourse relations used in the text plans are described in more detail in Section 4.

The translation methodology was first developed on a single Fable, "The Fox and the Grapes", until high coverage was achieved. The model was then tested on a set of 35 additional Fables from the DramaBank. Additional refining of the model was performed on a single story from PersonaBank, then tested on the remaining 107 blogs.

Gaps arise in the mapping from SIG to LSSTREE due to an individual annotator's personal choices when encoding their story and selecting WordNet and VerbNet propositions. Annotators may opt to select a proposition that paraphrases the original story in various ways. This allows for a range of story encodings, but may result in an unexpected LSSTREE or resulting realizations. Section 4 describes how the narrative sentence planner applies alterations to the LSSTREES in order to generate rich text that prioritize natural realizations. Thus the LSSTREES act as an intermediary representation during the course of the storytelling pipeline, sitting between story planning and sentence planning, and having access to the affordances of the entire pipeline.

3.3 Text Realization from Lexical-Semantic Story Trees

LSSTREES are the output of the translation process in Figure 3, from which natural language text can be generated using a surface realizer. The syntactic representation of LSSTREES are a one-to-one mapping to Deep Syntactic Structures (DSYNTS), the input to the real-time surface realizer, RealPro (Lavoie and Rambow, 1997), which is utilized at the final stage of generation in the Fabula Tales pipeline (seen later in Figure 8). RealPro handles morphology, agreement and function words to produce an output string. Gender, tense, co-reference, and articles are automatically handled by RealPro at generation time. The DSYNTS formalism distinguishes between arguments, modifiers, and between different types of arguments (subject, direct and indirect object etc.). Lexicalized nodes also contain a range of grammatical features used in generation. Figure 7 shows the DSYNTS representation for the LSSTREE in Figure 5. DSYNTS are ordered; the root is the main verb with required properties in XML format, including *lexeme* and *tense*. The *rel* argument indicates the relationship of the argument with respect to its parent. Nouns require an *article* argument, indicating a definite or indefinite article. Additionally, they can have a *gender* and *number*. Possession is represented structurally, so "the squirrel's reflection" is structured with "reflection" as the parent, and "squirrel" as the child, with the child also being possessive (*pro*).

Each lexeme from each LSSTREE node and information derived from the SIG are used to map the LSSTREE to DSYNTS. Text plans from the LSSTREES are created with the discourse relation

```
1  <dsyntnode id=1 class="verb" lexeme="leap" tense="past">
2    <dsyntnode article="def" lass="common_noun"
         gender="neut" lexeme="squirrel" number="sg"
         person="" rel="I"/>
3  </dsyntnode>
4  <dsyntnode id=2 class="verb" lexeme="see" tense="past">
5    <dsyntnode article="def" lass="common_noun"
         gender="neut" lexeme="squirrel" number="sg"
         person="" rel="I"/>
6    <dsyntnode article="def" lass="common_noun"
         gender="neut" lexeme="reflection" number="sg"
         person="" rel="II"/>
7  </dsyntnode>
8  <pdtbplan>
9    <relation name="contingency">
10     <proposition id="1" />
11     <proposition id="2" />
12   </relation></pdtbplan>
```

Figure 7: DSYNTS and text plan corresponding to the LSSTREE in Figure 5

Scheherazade Realization	LSSTREE Baseline, no Narrative Variation
A narrator placed a steely and large bowl on a back deck in order for a dog to drink the water of the bowl. A crazy squirrel approached the bowl. The second squirrel began to be startled because it saw the reflection of the squirrel. The squirrel leaped because it was startled and fell over the railing of the deck and because it leaped. The squirrel held the railing of the deck with a paw of the second squirrel. The squirrel fell, and the paw of the squirrel slipped off the railing of the deck.	The narrator placed the bowl on the deck in order for Benjamin to drink the bowl's water. The squirrel approached the bowl. The squirrel was startled because the squirrel saw the squirrel's reflection. The squirrel leaped because the squirrel was startled. The squirrel fell over the deck's railing because the squirrel leaped because the squirrel was startled. The squirrel held the deck's railing with the squirrel's paw. The squirrel's paw slipped off the deck's railing. The squirrel fell.

Table 7: The *Scheherazade* and LSSTREE baseline realizations of the *Startled Squirrel*

between DSYNTS nodes using the structure in Figure 7. Class properties are then written to a file, and the resulting file is processed by RealPro to generate the text.

Table 7 compares the *Scheherazade* realization (left-hand side) to the baseline realization as translated directly from the LSSTREES with no narrative variation (right-hand side). The baseline story is told in chronological order by a direct translation of the SIG timeline events into the surface order of the final realization. Discourse relations such as CONTINGENCY are always realized within a single sentence using *because* as a discourse cue.

The next section will evaluate the preservation of semantic content when mapping from *Scheherazade* to LSSTREES as measured by these baseline realization methods, in an effort to compare the ground-truth *fabula* of each story point without conflating the measure with the *sujet*.

3.4 Evaluating the Bridging Process

A prerequisite for producing stylistic variations of a story is the ability to generate a "correct" retelling of the story. To this end, we measure the semantic fidelity of our translation process, using text similarity metrics, as well as subjective measures of semantic similarity, grammaticality, and fluency. We compare the *Scheherazade* realization against the baseline generation by realizing LSSTREES without applying narrative variation as described in the previous section.

Pair #	*Scheherazade* Realization	LSSTREE Realization, no Narrative Variation
1	The squirrel leaped because it was startled and fell over the railing of the deck and because it leaped	The squirrel leapt because the squirrel was startled. The squirrel fell over the deck's railing because the squirrel leaped because the squirrel was startled.
2	The narrator greeted the woman and the acquaintance.	The narrator greeted Capt John and Ann.
3	The narrator didn't initially notice that the group of bugs had entered the apartment of the narrator.	The narrator did not initially notice that the bugs entered the narrator's apartment.
4	A group of persons began to dive around Great Barrier Reef, and the narrator entered some water.	The narrator entered the water.
5	The milkmaid began to plan for the milk to later transform into some cream, for the milkmaid to later make the cream into some butter and to later sell the butter at a market, for she to later buy some eggs, for a group of chickens to later hatch from the eggs, for the milkmaid to later sell it, to later buy a gown and for she to later wear it at a fairground, for every fellow to later admire the gown and to later court the milkmaid, and for the milkmaid to later shake the head of the milkmaid and to later ignore every fellow.	The milkmaid planned the milkmaid ignored every chap.

Table 8: Pairs of *Scheherazade* realizations and the LSSTREE baseline realizations

We evaluate the semantic fidelity of the realizations at the sentence level, rather than at the whole story level, for greater precision. We create an evaluation set of 320 blog pairs and 100 Fable pairs consisting of the *Scheherazade* realization and the equivalent baseline realization for each story point (Table 8 shows a subset of pairs). Sometimes this comparison results in a single sentence in *Scheherazade* being compared against more than one sentence in the baseline (e.g., pair #1 in Table 8). This is because two or more events that had been encoded in the SIG as taking place within a single story point have been split apart during LSSTREE creation and assigned a TEMPORAL discourse relationship. In addition, there are differences in how names are realized. In the baseline LSSTREE version in Table 7, the dog is named *Benjamin* whereas in the *Scheherazade* version, that character is known simply as *a dog* (also see pair #2). Pair #5 illustrates a case of an error in the process of creating the LSSTREE.

	BLEU	NIST	METEOR	ROUGE	STS
Fables	0.33	2.17	0.39	0.62	4.74
Blogs	0.29	1.98	0.38	0.61	4.54
All	0.30	2.02	0.39	0.61	4.59

Table 9: Metrics comparing *Scheherazade* and LSSTREE baseline realizations

We first apply a set of automated metrics that are commonly used to evaluate NLG output. We use the *Scheherazade* realization as the reference sentences against which the baseline realizations are evaluated, and apply the following metrics to each story pair using the e2e-metrics suite:[3] BLEU (Papineni et al., 2002), NIST (Doddington, 2002), METEOR (Denkowski and Lavie, 2014), and ROUGE-L (Lin, 2004) (first four columns of Table 9). As Table 9 shows, the results appear to be somewhat low: a BLEU score of 0.30 is much lower, for example, than the baseline system used in the E2E generation challenge. However, it is well known that these automatic metrics often do not reflect how well an NLG is actually performing (Belz and Reiter, 2006; Novikova et al., 2017).

We do not compare the original fables and blog stories to the *Scheherazade* and LSSTREE baselines in these evaluations. It is crucial to recall that these baselines do not have any sentence planning or narrative variation, and using these automated metrics would result in a biased comparison against the original story. Before applying sentence planning, we are primarily interested in verifying the semantic fidelity of the process of translating from the SIG to LSSTREES, rather than the generated language output. Section 4 conducts human subject evaluations that do include comparison of the original stories to those generated with our narrative sentence planning, which serves as a more fair comparison.

To measure the semantic fidelity of the translation, we conducted an evaluation using the semantic textual similarity metric (STS), obtained by human annotation,[4] with results shown in the final column of Table 9. Semantic textual similarity is defined as a scale that ranges from 1... 5, where 5 is *means exactly the same thing*, 4 is *means the same thing except for minor differences*, 3 is *means roughly the same thing*, 2 is *some important information is missing or different* (Cer et al., 2017). None of our pairs scored a 1, and only two scored a 2 where nested information from 2 fables were lost in translation (e.g., pair #5 in Table 8). Almost 80% of the pairs were scored a 5, and the human average is 4.59 over both datasets.

One decision made in the STS evaluation was to treat names as important semantic entities. As we explained above, during annotation, deictics like *I* are often annotated as *the narrator* and their realization can be changed during sentence planning, whereas *Scheherazade* does not realize names. In some cases, a character was given a name during annotation. Characters with names are common in the blogs, but in the Fables, characters are always known as their type *The Fox* or *The Wolf*. When treating names as important (i.e., less blog pairs are likely to be rated as a 5), we find a statistically significant difference between the STS means of blogs and fables that we hypothesize is due to name realization (paired t-test, df = 418, t = -2.78, $p < 0.01$). However, if names are treated as an unimportant part of the semantics of an utterance (i.e., more blog pairs are likely to be rated as a 5), the mean STS score of utterance pairs from the blogs are not statistically different from the mean STS score of the fables (paired t-test, df = 418, t = 0.065, p = 0.95).

Automatic metrics often conflate measures of fluency and naturalness with semantic correctness, and also penalize stylistic differences (Oraby et al., 2018). We therefore conducted an addi-

3. https://github.com/tuetschek/e2e-metrics
4. The annotator was an author of this article.

	Fluency		Grammaticality	
	Scheherazade	Baseline	*Scheherazade*	Baseline
Fables	3.63	3.22	4.08	3.54
Blogs	3.36	3.53	4.24	4.20
All	3.43	3.45	4.21	4.03

Table 10: Metrics comparing *Scheherazade* and LSSTREE baselines

tional evaluation for fluency and grammaticality with a human annotation task where we rate the *Scheherazade* and baseline realizations for each pair.[5] We used a Likert scale of 1 . . . 5 to state the degree of agreement with two statements: (1) The utterance is grammatical; and (2) The utterance is fluent and natural. The results for this evaluation are shown in Table 10, and show a high degree of grammaticality but a lower degree of fluency.

This bridge provides for a semantic mapping from SIG to LSSTREE while maintaining lexical information from the story points. These semantic similarity metrics measure the quality of the translation process, and the grammaticality and fluency metrics provide a rough estimate of the quality of the baseline LSSTREE realization prior to sentence planning. In subsequent evaluation with human subjects, we include these *Scheherazade* and LSSTREE baselines with the sentences generated with narrative sentence planning, as well as sentences from the original story texts, for a more comprehensive evaluation of stylistic expression. The application of these narrative variations are introduced in the next section.

4. Narrative Sentence Planning

The creation of the LSSTREES provide the syntactic and semantic grounding necessary to generate multiple *sujet* using narratologically inspired parameters. We develop a narrative sentence planner for Fabula Tales that takes the automatically generated LSSTREES and applies three narrative aspects that Lönneker (2005) describes: *Mood:Point of View*, *Mood:Distance*, and *Voice:Person* (Figure 8). We develop a sentence planner that implements parameters to model each of these narrative aspects, as we describe in detail below. Our sentence planner is based on the architecture of the sentence planner in the PERSONAGE NLG engine (Mairesse and Walker, 2008, 2011). We implement some of PERSONAGE's parameters related to *voice: person* and add new parameters that allow us to test particular narratologically inspired variations including *mood: distance* and *mood: point of view*. After manipulating the LSSTREES along these narrative dimensions, they are realized as text using DSYNTS and RealPro as described in Section 3.4.

We evaluate the sentence-level variations and show improvement over the baselines in the previous section. However, for some parameters, such as point of view and voice, it makes intuitive sense for this to be a story-wide decision, rather than a sentence-by-sentence decision. Thus, we design a story-level narratological structurer to ensure consistency in the generated styles and voices, as we discuss in Section 5.

5. We randomly mixed together the realizations so that the annotator would be blind to the source of the realization. The annotator was an author of this article.

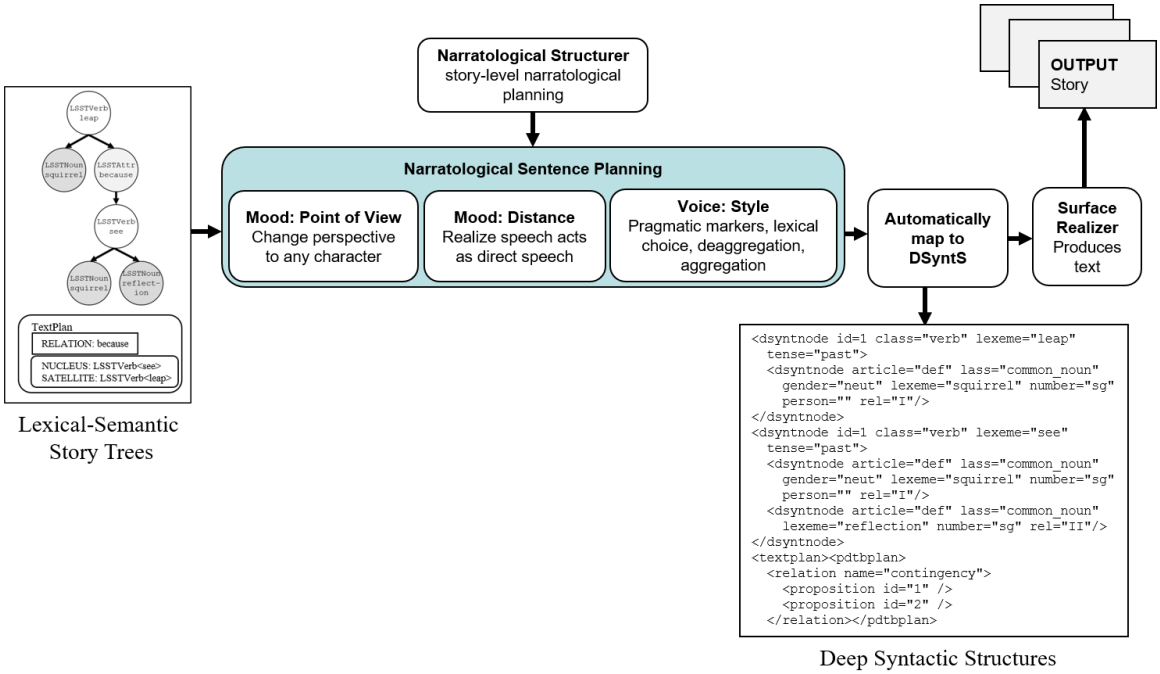


Figure 8: Fabula Tales pipeline applying narrative variations to LSSTREES

4.1 Mood: Point of View

Mood:Point of View is the "spatial, temporal, and ideological points of view from which events are described. Events can be described from the point of view of different characters." (Lönneker, 2005). Our architecture encodes characters, humanoid and non-humanoid, and props as unique actors or objects in the SIG that can be used for easy co-reference.

Biber (1991) claims that first person pronouns are markers of ego-involvement with a text. First person pronouns are often the subject of cognitive verbs, and indicate that the matter at hand is personal and an immediate mental interaction. In contrast to third person pronouns (and third person narration), first person pronouns create a different perspective in the narrative space, by restricting the perception of events to the eye of a particular character, and thus allows the audience limited perception, or focalization (Pizarro et al., 2003).

Any character in a story, including non-narrating or non-humanoid characters such as the squirrel in *Startled Squirrel*, can tell a story from their perspective using LSSTREES. Whenever first-person is desired, the LSSTREE sets this parameter when creating the DSYNTS mapping. A major advantage of the LSSTREES are that the deep linguistic representation allows for the specification of a change in point of view without manipulating the surface string or editing a template. Table 11 shows the DSYNTS for "the squirrel". In order to transform a sentence into the first person, from the DSYNT in Table 11, the `person` attribute is assigned to `1st` to specify a change of point of view to first person, reflected in Table 12. The RealPro surface realizer interprets the `person` attribute and automatically changes the lexeme present to "I". The LSSTREE representation tracks the identities of the characters and handles the realization of co-reference and possession (Lukin and Walker, 2015).

```
1 <dsyntnode article="def" class="common_noun" gender="neut"
      lexeme="squirrel" number="sg" person="" rel="I"/>
```

Table 11: DSYNTS for *The squirrel*

```
1 <dsyntnode article="def" class="common_noun" gender="neut"
      lexeme="squirrel" number="sg" person="1st" rel="I"/>
```

Table 12: DSYNTS for *I*

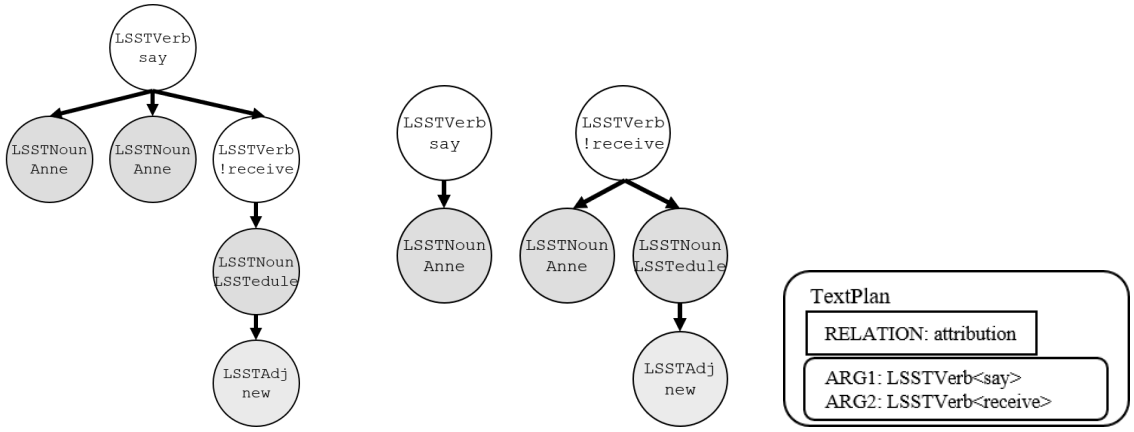
All the original texts from PersonaBank are told in the first person perspective, yet when they are annotated using the SIG there is no support for encoding different perspectives, because distinguishing between narrators is the job of the *sujet*, and not the *fabula*. To handle this, these stories are encoded with a "narrator" character, as we mentioned above. Just as the "squirrel" lexeme can be changed with a simple `person` attribute change, so too can the "narrator". In cases of multiple characters of the same type, e.g., two squirrel characters, the perspective change and subsequent co-reference tracking would only apply to the unique identifier each character is assigned during the LSSTREE creation as derived from the SIG.

4.2 Mood: Distance (Direct Speech)

Mood:Distance is the "combination of amount of information conveyed and narrator intrusion. Stereotypically, detailed information and low narrator participation indicate imitation or 'direct' dramatic mode, as opposed to a 'distant', mediated narrative mode. This parameter also affects the way in which speech is reproduced" (Lönneker, 2005). Our main manipulation of this variable is based on our supposition that the distance between the reader and the story can be altered by varying whether speech is direct or indirect.

When storytellers tell stories, they know what their characters are feeling, and can express it in the telling. Bal claims that "dialogue is a form in which the actors themselves, and not the primary narrator, utter language" (Bal, 1997) and that, in some cases, dialogue can make a narrative more dramatic. Speech acts in the SIG formalism are always encoded as indirect speech. In order to identify opportunities for direct speech (dialogue) the WordNet sense provided from the SIG and encoded in the LSSTREE is used to identify whether the main verb is a verb of communication. If so, the LSSTREE is broken apart into two separate trees: the utterance to be uttered, and the explanatory phrase, and are linked by the PDTB discourse relation of ATTRIBUTION (Prasad et al., 2008). There are many opportunities in both the DramaBank and PersonaBank to use direct speech: nine fables and forty eight blogs contained at least one speech act.

For example, in the sentence *Anne said she didn't receive the new schedule*, from the PersonaBank story called *Botched Training*, the verb *say* is identified as a verb of communication from VerbNet, with *Anne* as its subject (Figure 9a). The remainder of the tree starting from the verb "receive" as the root verb, which is what is to be uttered, is split it off from its parent verb of communication, resulting in two smaller trees (Figure 9b). After splitting, each tree is treated as a unique LSSTREE. A text plan is constructed consisting of the two LSSTREES linked by the ATTRIBUTION relation (Figure 9c). This text plan can then be realized in direct speech as *"I didn't receive the new schedule" Anne said.*



(a) LSSTREE for *Anne said she didn't receive the new schedule*

(b) Split LSSTREES for *Anne said* and *Anne didn't receive the new schedule*.

(c) Text Plan with the "Speech" discourse relation

Figure 9: LSSTREES and TextPlans for Direct Speech

4.3 Voice: Person

Voice:Person is the "narrator participation. A homodiegetic narrative instance is a character of the current narration (grammatical realization typically in the first person), while a heterodiegetic narrative instance is 'absent' from the current narrative and not referred to" (Lönneker, 2005). Different voices are showcased by combining different stylistic variations, including pragmatic marker insertion, lexical choice, and discourse structuring. To portray the *Voice:Person* parameter in Fabula Tales, we develop stylistic variations that, when combined, act as a character's speaking style.

Pragmatic Markers. Biber suggests that emotive, cognitive, modal and uncertainty words are indicators of personal stories, whereas these items are lacking in impersonal stories (Biber, 1991). Many of these are considered to be pragmatic markers, which we expect to be more prevalent in the natural language of the blogs. We define pragmatic markers following Biber for the following categories: acknowledgments (e.g., "oh"), emphasizers (e.g., "actually", "rather"), competence mitigations (e.g., "come on"), down tones (e.g., "I mean"), tag questions (e.g., "no?"), expletives (e.g., "damn"). We also emulate stuttering (e.g., "tr-trellis"), contractions, and exclamation insertions. Pragmatic marker insertion replicates PERSONAGE's mechanisms, which add nodes to the DSYNTS tree in the appropriate location (Mairesse and Walker, 2011). Table 14 lists a number of these pragmatic markers with a description and an example realization.

Lexical Choice. The WordNet and VerbNet senses from the SIG are used to manipulate the lexemes and structures of LSSTREES with synonym substitutions. Word senses are annotated when the SIG is created, as explained above, and preserved in the LSSTREE. Lexical choice can be controlled by implementing *word frequency* and *word length* as parameters, as in PERSONAGE. In one story from PersonaBank the narrator uses a comic book to try to kill some bugs that had been seen in his apartment. One of the sentences is *I smeared the bug's innards with the rolled comicbook*. The synset for "innards" contains 'viscera", "entrails", and "innards". Setting the *word frequency* parameter to be low could result in substituting "innards" with "viscera". Lexical substitutions for verbs are also possible but requires verifying that the synonym and its arguments

are interchangeable, for example, the verb "squash" in *I managed to squash the bug* is transformed to its argument equivalent "crush" in *I managed to crush the bug*.

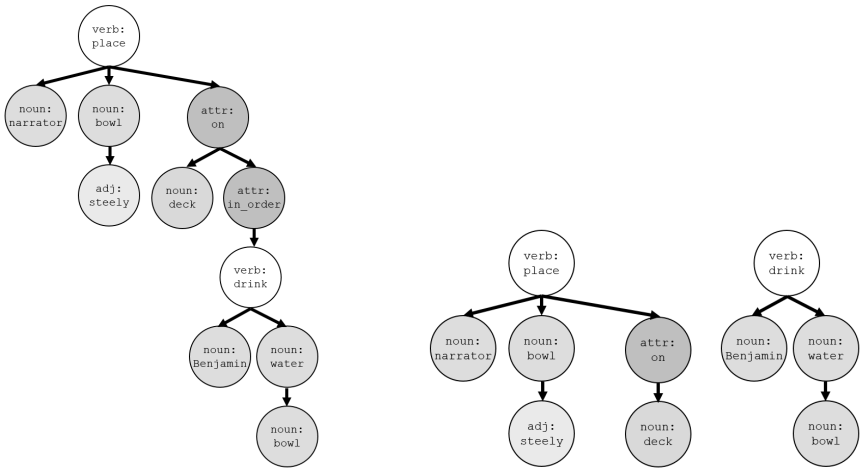
Although *word frequency* and *word length* are often inversely correlated, there are cases where short words are rare. Setting the *word length* parameter to be high for example could affect the lexical choice among the synset for "hue" in *The Fox and the Crow* where the Fox flatters the Crow by saying *The hue of her plumage exquisite*. Lexical substitutions that are available in the synset for "hue" include "chromaticity", which we see later in the variations in Table 20.

Deaggregation and Discourse Structuring. Traditionally, aggregation assumes an initial set of small semantic or syntactic pieces of information that can be readily combined. This may not always be the case; a clause may instead be packed with information that should first be decomposed, then restructured. We define *deaggregation* as the breaking apart or the decomposition of longer semantics into shorter semantics, which then affords the option and flexibility to aggregate the clauses in different arrangements, structures, or combinations.

We observe that when SIG semantics are converted to LSSTREES, some story points contain a sequence of nested predicates that, when realized without sentence planning intervention, result in particularly long sentences, for example:

> *The manager said she created the new schedule and the manager gave the new schedule to the employee in order for the employee to give the new schedule to Anne.*

The CONTINGENCY discourse relation is the target aggregation composition because these are highly likely to appear in narratives, and are abundant in PersonaBank (103 total instances). In the SIG, contingency clauses are expressed with the "in order to" relation. Story points with this relation are identified (Figure 10a) and split, to become the arguments of the CONTINGENCY relation as illustrated by the two distinct trees in Figure 10b.



(a) LSSTREE for *The narrator placed the steely bowl on the deck in order for Benjamin to drink the bowl's water.*

(b) Split LSSTREES for *The narrator placed the steely bowl on the deck* and *Benjamin drinks the bowl's water*

Figure 10: LSSTREES for deaggregation

Baseline	The narrator placed a steely and large bowl of water outside on the back deck **in order for** a dog to drink the water of the bowl.
Relations	CONTINGENCY (nuc:1, sat:2)
Content	1: put(narrator, bowl, deck) 2: dog(drink, bowl)
inOrder	I placed the bowl on the deck in order for Benjamin to drink the bowl's water.
becauseNS	I placed the bowl on the deck because Benjamin wanted to drink the bowl's water.
becauseSN	Because Benjamin wanted to drink the bowl's water, I placed the bowl on the deck.
NS	I placed the bowl on the deck. Benjamin wanted to drink the bowl's water.
N	I placed the bowl on the deck.
soSN	Benjamin wanted to drink the bowl's water, so I placed the bowl on the deck.

Table 13: Content assertions, texts plan, and possible realizations for CONTINGENCY

Table 13 shows new sentence planning variations for the CONTINGENCY relation. The *becauseNS* operation presents the nucleus, the primary clause (N) first, followed by a *because*, and then the satellite, the supporting clause (S). *becauseSN* and *soSN* reverse the order of the clauses. The nucleus and satellite can be treated as two different sentences (*NS*) or the satellite can be completely left off and only the nucleus realized (*N*). The richness of the discourse information present in the SIG enables the storytelling framework to implement additional discourse relations abundant in narratives in future work.

Defining Voice Models. Voice, in combination with changes to the point of view and direct speech, have the capacity to express the narrator as the storyteller, and the characters as speaking in their own style, similar to how other work has defined models using pragmatic markers to portray certain personality traits or character archetypes (Mairesse and Walker, 2008; Lin, 2016; Reed et al., 2011). Speech acts realized as direct speech can express a characters' particular style or way of speaking. An example of *Character Voice* is:

"Oh, well, I didn't receive the new schedule!" said Anne.

The acknowledgement "oh" and exclamation mark inside the direct speech reflect the mental and emotional state of the character as they express themselves. Compare this to *Direct Speech Only*:

"I didn't receive the new schedule", Anne said.

which makes use of direct speech, but not the stylistic variations of the voice. Finally, *Narrator Voice* is the voice of the primary narrator or storyteller which may use stylistic parameters, but excludes the direct speech:

Oh, Anne exclaimed that she did not receive the new schedule.

Voice models can be defined by setting all of the available parameters to have values between 0 and 1, in a similar way to how personality models were defined in the rule-based version of PERSONAGE (Mairesse and Walker, 2010a). Parameter values close to 1 indicate that the parameter should be used frequently, whereas parameter values near 0 indicate infrequent use of a parameter. In previous work, we build Laid-back and Shy models that are loosely based on the Extrovert and Introvert models from PERSONAGE (Mairesse and Walker, 2008; Rishes et al., 2013). Table 14

Model	Parameter	Description	Example
Shy voice	SOFTENER HEDGES	Insert syntactic elements (*sort of*, *kind of*, *somewhat*, *quite*, *around*, *rather*, *I think that*, *it seems that*, *it seems to me that*) to mitigate the strength of a proposition	*'It seems to me that he was hungry'*
	STUTTERING	Duplicate parts of a content word	*'The vine hung on the tr-trellis'*
	FILLED PAUSES	Insert syntactic elements expressing hesitancy (*I mean*, *err*, *mmhm*, *like*, *you know*)	*'Err... the fox jumped'*
Laid-back voice	EMPHASIZER HEDGES	Insert syntactic elements (*really*, *basically*, *actually*) to strengthen a proposition	*'The fox failed to get the group of grapes, alright?'*
	EXCLAMATION	Insert an exclamation mark	*'The group of grapes hung on the vine!'*
	EXPLETIVES	Insert a swear word	*'The fox was damn hungry'*

Table 14: Examples of pragmatic marker insertion parameters from PERSONAGE

shows the pragmatic markers used in combination to build each voice model. An additional, Neutral voice is constructed, which does not use any pragmatic markers, lexical substitutions, or aggregation constructions (equivalent to the LSSTREE baseline).

4.4 Evaluation of Sentence-Level Variations

We conduct a series of human evaluation tasks informed by previous research in this area (Callaway and Lester, 2002; Cheong and Young, 2008) to test the effectiveness of our narrative parameters on single sentences. Our evaluations measure the following narrative metrics:

- Narrative immediacy: to what degree is the reader engaged with the story and characters?
- Interest: to what degree would the reader desire to read the rest of the story?
- Correctness: to what degree is the narrative well-formed?
- Preference: which framings do readers generally prefer to read?

We hypothesize that generating stories by varying point of view (H_1), character or narrator voice (H_2), and aggregation operations (H_3), will have an effect on these narrative metrics.

4.4.1 POINT OF VIEW AND VOICE: ENGAGEMENT AND INTEREST

We examine how point of view and voice interact with engagement and interest in a single sentence from a story, and hypothesize that excerpts told in different points of view and voice will have an effect on engagement and interest (H_1 and H_2). Native English speakers on Mechanical Turk were presented with a one sentence summary of one of seven stories from PersonaBank and six generated variations of one sentence from that story. These sentences are framed as "possible excerpts that could come from this summary". Table 15 shows an example of the *Embarrassed Teacher* story from PersonaBank, the summary, and its six retellings. Narrative variations include the first person with a neutral, shy, and laid-back voice, and a third person with a neutral voice, as described in

Summary A teacher's slip fell down in the middle of teaching a class.	
Source	Example
Original	Nervously I looked down to see that my underslip had somehow made its way to the floor.
Scheherazade	The narrator noticed that the ankle of the narrator was observed.
3rd neutral	The narrator noticed for the narrator's ankle to be observed.
1st out	Oh I noticed for my ankle to be damn observed!
1st neutral	I noticed for my ankle to be observed.
1st shy	I noticed for my ankle to be so-somewhat observed.

Table 15: Variations presented to Turkers for interest and narrative immediacy

Section 4.3. Additionally, an excerpt from the original story from which the generated stories were derived is compared to test how close the best narrative sentence planning realization comes to matching the natural language of the blog. The strictly template-based *Scheherazade* realization was also included. Subjects rate each excerpt on a 1 . . . 5 point scale for their interest in wanting to read more of the story based on the style and information given in the excerpt, and to indicate their engagement with the story, given the excerpt (Lukin and Walker, 2015).

We performed a set of ANOVAs designed with the repeated items as categorical, independent variables, i.e., style (view and voice pairs) and story content (the particular sentence in question), subjects as a random variable, and aggregated across multiple items.[6] Style has an effect on interest ($F(1) = 204.08$, $p < 0.0001$), as does story content ($F(9) = 7.32$, $p < 0.0001$), but there is no interaction between style and story content. This may be interpreted as: style affects the sentence, and there is a random effect of story content, but interest preference is independent of the style and story content.

Style similarly has an effect on engagement ($F(1) = 224.24$, $p < 0.0001$) and story content ($F(9) = 5.49$, $p < 0.0001$). However, there is an interaction between style and story content ($F(9) = 1.65$, $p < 0.1$), which suggests that for engagement, but not for interest, certain styles of narration are more appropriate or preferred than others given the context of the story. For example, subjects comment that the "curse words are used to express the severity of the situation wisely" and "adding the feeling of nervousness and where she looked made sense", acknowledging the style fitting the situation. Information from the story may be used to influence and produce a more engaging realization. We briefly discuss how being cognizant of content can influence realization in future work.

Table 16 shows the means and standard deviation for engagement and interest for each combination of voice and person. An ordered ranking emerges for both engagement and interest: the original sentence from the blog is scored highest, followed by first-person laid-back, first-person neutral, first-person shy, *Scheherazade*, and third-person neutral.

For the subsequent analyses, the key independent variable was point of view and voice pairs (i.e., style). Bonferroni correction was applied to paired t-tests of style on engagement (full results in Table 29). For engagement, there are statistically significant differences between the following styles in the ordered list: original and first laid-back, first neutral and first shy, and first shy and *Scheherazade*. However, there are no other differences between sentences. Therefore, we observe

6. A linear effects model was not used because our item independent variables are not mixed.

Style	Engagement		Interest	
	Mean	Std err	Mean	Std err
Original	3.98	(1.07)	3.91	(0.99)
1st-laid-back	3.27†	(1.39)	3.02†	(1.21)
1st-neutr	3.00	(1.19)	3.02	(1.37)
1st-shy	2.73†	(1.26)	2.81†	(1.27)
Scheherazade	1.95†	(1.07)	1.90†	(1.05)
3rd-neutr	1.93	(1.06)	1.87	(1.01)

Table 16: Means and standard deviation for engagement and interest in perceptions experiment (higher is better; † indicates statistical significance between the marked style and the style in the row above; see Tables 29 and 30 for detail)

that H_1 is supported, with statistically significant differences in point of view realizations for the engagement metric, as well as H_2, with statistically significant differences in voice between laid-back, shy, and neutral.

Similarly, Bonferroni correction was applied to paired t-tests on style and interest (full results in Table 30). Results for interest follow a similar trend, showing a statistically significant difference between original and first laid-back, first neutral and first shy, and first shy and *Scheherazade*, again supporting H_1 and H_2 for the interest metric. There are no other differences between ordered pairs.

4.4.2 DISCOURSE STRUCTURING: CORRECTNESS AND PREFERENCE

We examine how discourse structuring interacts with correctness and preference in a single sentence from a story. We hypothesize that the deaggregation and discourse structuring variations will effect reader preferences and belief about the correctness of the narrative (H_3). We explore (1) how the variations compare to each other; (2) if they come close to the natural language of the original blog story; and (3) if the narrative sentence planning realization surpasses the *Scheherazade* realization. We create a Mechanical Turk experiment showing an excerpt from the original story, where we tell our qualified Turkers that "any of the following sentences could come next in the story" (Table 17). Subjects are queried about the variations in terms of correctness and goodness of fit within the story context. They are then asked to rank the sentences by personal preference (in experiment 1, we showed 7 variations where 1 is best, 7 is worst; in experiment 2 we showed 3 variations where 1 is best, 3 is worst). We emphasize in the prompt that subjects should read each variation in the context of the entire story, and encourage them to reread the story with each new sentence to understand this context (Lukin et al., 2015).

We performed a set of ANOVAs designed with the repeated items as categorical, independent variables, i.e., realization (variation) and story content (the particular sentence in question), subjects as a random variable, and aggregated across multiple items.[7] In the first experiment, seven native English speakers on Mechanical Turk analyzed 16 story segments from different blogs in PersonaBank with the following variations: the original story, soSN, becauseNS, becauseSN, NS, N, and the non-deaggregated realization. As expected, realization had an effect on correctness ($F(6) = 9.8$, $p < 0.0001$) and preference ($F(6) = 31.7$, $p < 0.0001$) supporting hypothesis H_3 that the realizations are distinct from each other and there are preferences among them, as well as varying degrees of

7. A linear effects model was not used because our item independent variables are not mixed.

Story	
This is one of those times I wish I had a digital camera. We keep a large stainless steel bowl of water outside on the back deck for Benjamin to drink out of when he's playing ouside. His bowl has become a very popular site. Throughout the day many birds drink out of it and bathe in it.	
Source	Example
Original	The birds literally line up on the railing and wait their turn.
Scheherazade	The birds organized themselves on the deck's railing.
becauseSN	Because the birds wanted to wait, they organized themselves on the deck's railing.
becauseNS	The birds organized themselves on the deck's railing because the birds wanted to wait.
soSN	The birds wanted to wait, so they organized themselves on the deck's railing.
None	The birds organized themselves on the deck's railing in order for the birds to wait.
NS	The birds organized themselves on the deck's railing. The birds wanted to wait.
N	The birds organized themselves on the deck's railing.

Table 17: Deaggregation and Discourse Structuring Variations presented to Turkers for correctness and preference judgments

Realizations	Correctness		Preference	
	Mean	Std err	Mean	Std err
Original	1.83	(1.34)	2.38	(2.28)
soSN	2.32†	(1.26)	3.07†	(1.89)
becauseNS	2.44	(1.28)	3.65†	(1.78)
becauseSN	2.45†	(1.26)	3.73	(1.93)
NS	2.69†	(1.13)	4.25†	(1.53)
None	2.72	(1.10)	4.86†	(1.72)
N	3.01	(1.14)	4.90	(1.47)

Table 18: Means for correctness and preference for discourse structure experiment 1 (lower is better; † indicates statistical significance between the marked realization and the realization in the row above; see Tables 31 and 32 for detail)

grammaticality. Story content had no effect on correctness or preference, suggesting that all stories were well-formed and there were no outliers in the story selection. We find an interaction between realization and story content for correctness ($F(2, 110) = 1.83$, $p < 0.0001$) and preference ($F(2, 110) = 3.24$, $p < 0.0001$), thus subjects' preference of the realization are based on the context of the story, unlike in the previous analysis of point of view for engagement and interest.

Table 18 shows the means and standard deviations for correctness and preference rankings for each realization in the first experiment. Averaged across all stories, there is a clear order for correctness and preference: Original, soSN, becauseNS, becauseSN, NS, non-deaggregated (indicated as None), and N.

For the subsequent analysis, the independent variable was if deaggregation was performed, and with which aggregation discourse structure construction. Bonferroni correction was applied to paired t-tests of realization on correctness (full results in Table 31). There are statistically significant difference in correctness between Original and soSN, and between becauseNS and becauseSN, as well as becauseSN and NS. Similarly, Bonferroni correction was applied to paired t-tests of real-

Realizations	Correctness		Preference	
	Mean	Std err	Mean	Std err
Original	1.57	(0.92)	1.37	(0.59)
soSN	2.49†	(1.29)	1.93†	(0.67)
Scheherazade	3.50†	(1.43)	2.70†	(0.57)

Table 19: Means for correctness and preference for discourse structure experiment 2 (lower is better; † indicates statistical significance between the marked realization and the realization in the row above; see Tables 33 and 34 for detail)

ization on preference (full results in Table 32). Results for preference follow a similar trend, with the addition of soSN and becauseNS.

These results indicate that the original sentence is the most correct and preferred. In a qualitative evaluation, subjects commented that while all variations were sufficient, most were "boring", except for the original blog story excerpt. The N and NS variations are overall ranked the lowest because they sometimes produce stilted language and remove pieces of content. However, in a few instances, these variations are ranked highly because the information they remove was deemed to be redundant in text realization or repeated content, which we posit shows support for the interaction between realization and story content.

In a second experiment, we compare the original blog sentence with the highest scoring discourse structure variation with a point of view change, and the realization produced by *Scheherazade*. We expect that *Scheherazade* will score poorly in this instance because it cannot realize deictic expressions to change point of view from third person to first person, even though it is derived directly from the SIG representation. Seven native English speaking subjects analyzed each of the 19 story segments in a similar experimental setup as the deaggregation experiment 1.

Our ANOVAs were conducted following the same design as the first experiment in this section. Realization had an effect on correctness ($F(2) = 6.78$, $p < 0.0001$) and preference ($F(2) = 131.9$, $p < 0.0001$), again, supporting hypothesis H_3. Story content had no effect, suggesting that there were no outlier stories, and there was an interaction between realization and story content for correctness ($F(2, 47) = 5.48$, $p < 0.0001$) and preference ($F(2, 47) = 9.25$, $p < 0.0001$), suggesting that subjects' evaluation is based on the realization and the context of the story. Table 19 shows the means and standard deviations for correctness and preference rankings for the realizations in the second experiment. There is a clear order for correctness and preference: original, soSN, *Scheherazade*.

Bonferroni correction was applied to paired t-tests of realization on correctness and preference (full results in Tables 33 and 34). For the majority of the stories, subjects do not select *Scheherazade* because of "the narrator" realization, commenting "forget the narrator sentence. From here on out it's always the worst!". However there are three story segments where *Scheherazade* is rated on average higher than soSN. Upon closer examination, these story segments do not contain "I" or "the narrator" in the story content, so the sentence is evaluated without the "narrator" bias. However, even without that bias, soSN still outranks *Scheherazade*: in a story about a protest at the G20 summit, the soSN realization:

The leaders wanted to talk, so they met near the workplace.

is much more natural than the *Scheherazade* realization:

> *The group of leaders was meeting in order to talk about running a group of countries and near a workplace.*

These evaluations have shown that realizations using the first person point of view, pragmatic features, and aggregation variations, are more engaging, interesting, correct, and preferred than the *Scheherazade* baselines and LSSTREE baselines without narrative. In the next sections, we explore how to intelligently plan and extend the narrative sentence planner to support story-level variations.

5. Narratological Structurer

The previous sections have shown that Fabula Tales' sentence planning parameters are capable of generating hundreds of sentences for any story that has first been annotated as a SIG, with high semantic fidelity to the *fabula* of the original story. Table 20 shows complete story generation achieved by setting several parameters for *The Fox and the Crow*, both using direct speech and different voice models for each character. These outputs are generated with our default story-level narratological structurer: this simply sets the parameters for the whole story to be consistent, e.g. if the DIRECT SPEECH parameter is set to 1, direct speech will be used throughout the story, rather than alternating between direct and indirect.

Variation 1: Shy Crow and Laid-Back Fox	Variation 2: Laid-Back Crow and Shy Fox
The crow sat on the tree's branch. The crow thought "I will eat the cheese on the branch of the tree because the clarity of the sky is somewhat beautiful." The fox observed the crow. The fox thought "I will obtain the cheese from the crow's nib." The fox averred "I see you!" The fox alleged "your beauty is quite incomparable, okay?" The fox alleged "your feather's chromaticity is exquisite." The fox said "if your voice's pleasantness is equal to your visual aspect's loveliness you undoubtedly are every birds' queen!" The crow thought "the fox was somewhat flattering." The crow thought "I will demonstrate my voice." The crow loudly cawed. The cheese fell. The fox snatched the cheese. The fox said "you are somewhat able to sing, alright?" The fox alleged "you need wits!"	The crow sat on the tree's branch. The crow thought "I will eat the cheese on the tree's branch because the sky's limpidity is beautiful". The fox observed the crow. The fox thought "I will obtain the cheese from the crow's pecker." The fox averred "I see the bird." The fox alleged "your beauty is somewhat incomparable." The fox alleged "your feather's chromaticity is somewhat exquisite." The fox said "if your voice's sweetness is somewhat equal to your appearance's beauteousness you undoubtedly are every birds' queen." The crow thought "the fox was flattering, you know, okay?" The crow thought "I will demonstrate my voice." The crow loudly cawed. The cheese fell. The fox snatched the cheese. The fox said "you are somewhat able to sing." The fox alleged "you need wits."

Table 20: *The Fox and the Crow* variations produced by the Narratological Structurer

Now, however, we consider that there is no guarantee that a naïve combination of all these narrative parameters will produce an appropriate narrative flow. First of all, it is clear that narrative text generation at the story-level should maintain a degree of sentence-by-sentence consistency. For example, the character or narrator voice and person parameters should not dramatically change mid-story without reason. Similarly, point of view should remain consistent throughout a story segment.[8] However, other narrative aspects, such as the use of direct speech and different syntactic constructions, may produce better stories if they are varied throughout a story.

8. In a story with multiple chapters or segments, the style or point of view may change between segments, but within a segment this is generally consistent.

We design and build a narratological structurer for Fabula Tales that sits above the sentence planner and dictates narrative operations to the sentence planner (see Figure 8). To train the narratological structurer, we undergo two phases: overgenerate and rank. In the **Overgenerate Phase**, we naïvely plug the sentence-level parameters into sentences to generate an abundance of training data (Section 5.1). We design a Create-Your-Own-Story paradigm for the **Rank Phase**, allowing subjects to construct a story sentence-by-sentence by selecting from the sentences generated in the Overgenerate phase (Section 5.2). Subject choose sentences that best contribute to the overall narrative flow. The rankings measure the effectiveness of each narrative parameter in the selected sentences and are utilized by the narratological structurer to make story-level generation decisions.

We evaluate the narratological structurer's generation capabilities and test exploratory hypotheses based on narrative theories from Lönneker (2005) and observations from Biber (1991). Similar to as before, we hypothesize that generating stories by varying point of view (H_1), character or narrator voice (H_2), and aggregation operations (H_3), will have an effect on reader perceptions. We add a new hypothesis (H_{2a}) that direct speech in isolation will have an effect on reader perceptions (Section 5.3). Finally, we conduct a classification exercise to determine if the data and features collected from the Overgenerate and Rank phases can be used to identify which pre-generated sentences will be the most preferred (Section 5.4).

5.1 Overgenerate: Generating Training Data

Depending on the narratological, structural, or lexical features present in the encoding, Fabula Tales produces different variations when generating variations. For this study, four stories from PersonaBank are used, each seven sentences in length. A total of 2330 different sentences variations were generated from the original 28 baseline sentences. Sentences are generated with combinations of all the parameters discussed in Section 4: point of view, direct speech, and voice.

#	*Botched Training* Variations
1	I rather excitedly entered PF Changs because the manager wanted to train me
2	I excitedly entered PF Changs in order for the manager to train me
3	The manager wanted to train me, so I excitedly entered PF Changs
4	Ok, I excitedly entered PF Changs in order for the manager to train me, right?
5	Because the manager wanted to train me, I excitedly entered PF Changs
6	The manager wanted to train Anne, so she excitedly entered PF Changs, as it were
7	Because the manager wanted to train Anne, she excitedly entered PF Changs!!
8	Anne excitedly entered PF Changs
9	Essentially, ok, the manager wanted to train Anne, so she excitedly entered PF Changs
10	Actually, Anne excitedly entered PF Changs in order for the manager to train her
11	The director wanted to train Anne, so she excitedly entered PF Changs
12	The manager wanted to train me, so I excitedly entered PF Changs, okay?

Table 21: Variations of first sentence of *Botched Training* Story

Table 21 illustrates a subset of these variations of the first sentence from the *Botched Training* story from PersonaBank. The sentence can be deaggregated into two clauses: *Anne excitedly entered PF Changs* and *The manager wanted to train Anne*. Outputs 1, 3, 5, and 8 have different discourse constructions, including the arguments, using different discourse cues, or removing the less important argument completely on the assumption that it is likely to be redundant. Outputs 2 and 4 do not deaggregate, and instead realize the most straightforward logical form. Output 1 has

the emphasizer "rather", outputs 4 and 9 have the acknowledgement "ok", and output 7 has exclamation marks. There are variations both in the first and third person point of view. Because there is no speech act in the sample sentence in the Table 21, this utterance is defined as evoking "Narrator Voice".

5.2 Rank: Training the Narratological Structurer

The goal of the Rank phase is to learn story-level parameters for training a narratological structurer that preserves the desirable properties of random, probabilistic generation of different story versions, while at the same time making sure that these parameters take feedback from readers into consideration.

The Create-Your-Own-Story paradigm allows subjects to build a story sentence-by-sentence by selecting from a subset of the overgenerated sentences. Figure 11 shows the experimental design. For tractability, we downselect the generated sentences into a set of 5 variations per sentence from the Overgeneration phase, resulting in 28 sentences per story, yielding a total of 140 sentences from which the subject can create use to create stories. The subset was designed to showcase each feature that can appear in at least one sentence in a variety of combinations with other features. Subjects select the sentences they like best. At the bottom of the experiment, the progression of the reconstructed story is dynamically updated so subjects can read the full story to see how it flows. Subjects are encouraged to read each sentence within the context of the entire reconstructed story. At any time, they may select a different sentence, yielding a dynamic update to the reconstructed story. When finished, the subjects rate how much they like their story on a 5-point Likert scale, and then are asked to give detailed feedback about why they selected the sentences they did.

Nine subjects on Mechanical Turk who were prequalified for language-based reading and comprehension tasks completed this task. A total of thirty reconstructed stories were created with an average enjoyment score of 3. Showing the reconstructed story at the bottom of the experiment allowed subjects to engage with their own perceptions of the flow of the narrative. Many annotators commented about the flow of the story and keeping consistency.

We make two notable observations from the qualitative feedback: (a) annotators tried to create stories with a good flow and consistency; and (b) pragmatic marker features as a way to create a character voice are pragmatically odd in many cases. This qualitative and quantitative analysis gives us insight into which features are selected, how often, and how the narratological structurer can use the high ranked features in generating future stories. Deaggregation, direct speech, and contractions are popular and used more than 50% throughout a story. However, some pragmatic markers for character voice in direct speech or narrator voice are not used as consistently because they are pragmatically odd, and do not take context into consideration at the individual sentence-level or across the entire story.

By analyzing the sentences that were selected and those that were not selected by annotators during the "Create Your Own Story" experiment, we design two metrics for training the narratological structurer's parameterizable model: the $catRatio$ metric is aimed at learning the appropriateness and placement of the narrative features in individual sentences, and the $perStoryRatio$ discovers the balance of how many times a particular feature should occur within a story. In the next section, these statistics are used to test whether the application of the $catRatio$ and $perStoryRatio$ improves the quality of the story-level generation.

Construct Your Own Story

Sentence Set 1:

● I rather excitedly entered PF Changs because the manager wanted to train me.
○ The manager wanted to train me, so I excitedly entered PF Changs, okay?
○ Really I excitedly entered PF Changs in order for the manager to train me.
○ Because the manager wanted to train me, I excitedly entered PF Changs, as it were.
○ I excitedly entered PF Changs in order for the manager to train me.

.
.
.

Sentence Set 7:

○ ``I see, yeah, I didn`t receive the new schedule!``, I said.
○ ``I didn`t receive the new schedule!``, I said. Oh my God, right?
○ I said I didn`t receive the new schedule, actually
○ ``Yeah, I see, I didn`t receive the new schedule!``, I said, you know.
● ``Oh yeah, I didn`t receive the new schedule!``, I said.

Your Reconstructed Story:

1. I rather excitedly entered PF Changs because the manager wanted to train me.
2. ``Yesterday I scheduled you in order for me to train you and you didn`t show up``, the manager lazily said.
3. Because the schedule demonstrated my punctuality, I was very confused.
4. ``The schedule was erroneous``, the disgruntled manager lazily said.
5. I insistently questioned the manager because the manager lazily said the schedule was erroneous.
6. ``I created a new schedule and I gave the new schedule to the employee in order for her to give the new schedule to you``, the manager stated
7. ``Oh yeah, I didn`t receive the new schedule!``, I said.

Figure 11: Create-Your-Own-Story experimental design (sentence sets 2-6 omitted for space)

We define the **Category-Ratio** metric, the percentage of the time a particular feature i is used with respect to the other features in each category, as:

$$catRatio_i = \frac{sel_i}{sel_{i.total}} \tag{1}$$

where i is an item in a feature category, sel_i is the number of times a sentence with feature i was selected, and $sel_{i.total}$ is the total number of selected sentences in that feature category. For example, from Table 22, there are a total of 217 sentences that were selected by the subjects that had the potential for an acknowledgment to be inserted ($sel_{ack.total}$ = 217). Of those 217 selected sentences, only 29 actually had an acknowledgement present ($sel_{ack.present}$) while the majority, 188, did not have the acknowledgment realized ($sel_{ack.\neg present}$). Thus, $catRatio_{ack.present} = \frac{29}{217} = 0.13$, indicating 13% of the sentences selected contained an acknowledgement, whereas, $catRatio_{ack.\neg present} = 0.87$ indicates that the other 87% of the selected sentences did not have the acknowledgement.

Many of the individual voice parameters rarely or never appear in the selected sentences, including "I mean", "come on", "like", and "no?". We believe placement is the problem because these stories were not generated with any story-wide constraints; indeed, this is what we aim to learn through this experiment. Popular features were the acknowledgement "yeah" and the emphasizers "really", "very", and "actually". The competence mitigation toner is rarely used, and the tag question category is never selected. Contractions were selected an overwhelming 85%, and may be more likely to emulate the flow and naturalness of everyday speech, regardless of narration or direct speech. For sentences with exclamations, $catRatio_{exclam.present}$ is 40%. Table 22 also shows that

Category	Features	sel	$catRatio$
Acknowledgment	yeah	14	.06
	right	4	.02
	{I see, oh}	3	.01
	{oh my god, ok, well, oh yeah}	1	$<$ .01
	{great, okay?}	0	0
	$ack.present$	29	.13
	$ack.\neg present$	188	.87
	$ack.total$	217	-
Competence mitigations	obviously	3	.01
	come on	0	0
	$mit.present$	3	.01
	$mit.\neg present$	214	.99
	$mit.total$	217	-
Down	rather	17	.08
	{I mean, like, somewhat}	0	0
	$down.present$	17	.08
	$down.\neg present$	200	.92
	$down.total$	217	-
Emphasizers	actually	11	.05
	really	9	.04
	{great, very, you know, especially}	1	$<$ .01
	{as it were, basically, essentially, obviously}	0	0
	$emp.present$	26	.12
	$emp.\neg present$	191	.88
	$emp.total$	217	-
Exclamation	$exclam.present$	50	.40
	$exclam.\neg present$	74	.60
	$exclam.total$	124	-
Contraction	$contr.present$	29	.85
	$contr.\neg present$	5	.15
	$contr.total$	34	-
Discourse Structuring	$disc.present$	67	.74
	$disc.\neg present$	23	.26
	$disc.total$	90	-
Direct speech	$ds.present$	40	.69
	$ds.\neg present$	18	.31
	$ds.total$	58	-

Table 22: Feature categories for ranking generated stories

74% of selected sentences have a discourse structure variant. There is also a slight preference for direct speech, with a $catRatio_{ds.present}$ of 69%.

The $perStoryRatio$ is defined for for pragmatic markers based on the observed data such that 60% of reconstructed stories do not have any voice or style features, 27% have only one, 7% have two, and none have more than two. The same pragmatic feature is never selected twice in a story. While a few pragmatic features may be good for expressing character voice, too many repetitions of the same or similar markers appear to be perceived as unnatural.

Generated Variant 1	Generated Variant 2
Because the manager wanted to train Anne, she excitedly entered PF Changs!! The manager lazily said yesterday she scheduled Anne in order for her to train Anne and Anne didn't show up. Anne was confused. The rather disgruntled manager lazily said the schedule was erroneous. Right, the manager lazily said the schedule was erroneous, so Anne insistently questioned the manager. "I created a new schedule and I gave the new schedule to an employee in order for her to give the new schedule to you", the manager melodramatically said. "Oh I see, I didn't receive the new schedule!", Anne said.	I excitedly entered PF Changs in order for the manager to train me. "Yesterday I scheduled you in order for me to train you and you didn't show up", the manager lazily said. I mean, I was confused because the schedule demonstrated my punctuality. "The schedule was erroneous", the disgruntled manager lazily said. I insistently questioned the manager because the manager lazily said the schedule was erroneous. "I created a new schedule and I gave the new schedule to an employee in order for her to give the new schedule to you", the manager stated. "I didn't receive the new schedule!", I said. Oh my God, right?

Table 23: Two stories constructed by the trained Narratological Structurer

The narratological structurer uses the $catRatio$ and $perStoryRatio$ metrics to dictate to the sentence planner when to apply a particular narrative parameter to the LSSTREES when generating the fully realized stories. For each narrative parameter, the narratological structurer uses a probability distribution to determines what value to assign to that parameter according to the $catRatio$ percentages. It also keeps track of how many of each parameter have been used within the story so far, if applicable, and makes generation decisions according to the $perStoryRatio$. In the next section, we test the effectiveness of generating full stories using the learned $catRatio$ and $perstoryRatio$ statistics in the narratological structurer.

5.3 Evaluation of Narratological Structurer

Stories generated by the trained narratological structurer are shown in Table 23. Variant 1 is told in the third person perspective, with a variety of sentence constructions. Because this story is told in the third person, the use of direct speech gives opportunities to the characters to express themselves in their own voice, e.g. "Oh I see" in the direct speech of Anne. Variant 2 is told entirely in the first person perspective of the employee, and all direct speech, plus narrator observations, such as "Oh my God, right?", that cannot appear in a third person narration.

We conduct three evaluations of the narratological structurer. The first compares stories generated according to these statistics against the LSSTREE baseline presented in Section 3.4 (Rank vs. Baseline). These tests consist of a number of ablation tests to study the narrative hypotheses H_1 - H_3. The second compares the stories generated by the narratological structurer to stories generated with randomly assigned narrative parameter values (Rank vs. Random). Finally, we compare the randomly generated stories to the LSSTREE baselines (Random vs. Baseline).

Rank vs. Baseline. We use the narratological structurer to generate a set of eleven stories from PersonaBank in the following manner: a subset are generated in the first person point of view but with no other parameters (to be compared to the third person baseline, to test H_1); a subset are generated with direct speech according to the narratological structurer (to be compared to the indirect speech baseline, to test H_{2a}); a subset are generated with voice parameters according to the narratological structurer (to be compared to the neutral voice baseline, to test H_2); and a subset are generated with deaggregation and discourse structuring according to the narratological structurer (to

be compared to the no deaggregation baseline, to test H_3). Each pair of Rank-Baseline stories are annotated by seven subjects on Mechanical Turk, who were asked which story they prefer (framed as "story A or story B"). The results indicated a preference for the stories generated by the $ratio$ statistics over the baseline stories 75% of the time.

We conducted the subsequent analyses using an ANOVA with independent variables of point of view (H_1), direct speech (H_{2a}), direct speech and style (H_2), and deaggregation and discourse structuring (H_3). The dependent variable was preference.

$\mathbf{H}_1$ claims that stories told in the first or third person point of view will effect reader preferences. Point of view is not statistically significant, nor is there an effect or interaction of story content. One subject observed in qualitative feedback that first person can lead to more opportunities for the characters to describe their feelings, while another identified that the first person story added a "sense of immediacy" and "tension". However, we find that other subjects "simply preferred the third person" without providing additional rational.

$\mathbf{H}_{2a}$ claims that direct speech will effect reader preferences, and we observe a statistically significant difference in preference between stories generated by the narratological structurer using direct speech and baseline stories ((1, N=42) = 23.3, p $<$ 0.0001). No effect of story content or interaction is observed, so this finding is independent of topic of the story. Readers commented that alternating narrative description in the non-speech sentences and dialogue adds personality to the story.

$\mathbf{H}_2$ claims that pragmatic markers, when used as character or narrator voice and generated according to the $ratio$ statistics, will effect reader preferences. We observe a statistically significant difference in preference between stories generated by the narratological structurer using voice and baseline stories ((1, N=98) = 74.0, p $<$ 0.0001). While there is no effect of story content alone, there is an interaction between pragmatic markers and story content ($F(5, 98) = 17.5, p < 0.0001$) suggesting that different preferences for pragmatic markers emerge in different contexts.

We examine in greater detail the direct speech and pragmatic marker interaction with the generated texts similar to those generated in Table 24. First, we compare *Direct Speech Only* and *Narrator Voice*. *Direct Speech Only* is preferred 100% of the time. Subjects say *Direct Speech Only* is easier to understand, which is conceivable because the direct speech breaks up the narrative into alternating speech acts and narration segments. When comparing stories with a *Narrator Voice* to stories with *Character Voice*, *Character Voice* stories are preferred 95% of the time.

Baseline	The manager said the schedule was erroneous. Anne questioned the manager because the manager said the schedule was erroneous. ... Anne said she didn't receive the new schedule.
Direct Speech Only	"The schedule was erroneous", the manager said. Anne questioned the manager because the manager said the schedule was erroneous. ... "I didn't receive the new schedule", Anne said.
Narrator Voice	The manager said the schedule was obviously, erroneous. Anne questioned the manager because the manager said the schedule was erroneous. ... Yeah, Anne said she didn't receive the new schedule.
Character Voice	"The schedule was obviously, erroneous", the manager said. Anne questioned the manager because the manager said the schedule was erroneous. ... "Yeah, right, I didn't receive the new schedule", Anne said.

Table 24: Excerpts of different speech conditions in rank vs. baseline ablation test

However, comparing *Character Voice* stories with *Direct Speech Only* stories, we observe a moderately strong preference (71%) for *Direct Speech Only*. One subject comments that *Character Voice* with the pragmatic marker voice features yields "mixed results". Subjects comment on what might be an issue of context insensitivity. For example, in *Character Voice*, the generated text includes *Yeah, right* as part of what Anne says in the direct speech. Subjects identified that this creates a tone they believe was not appropriate with respect to Anne's tone in the rest of the story. While these voice and style features are more acceptable as *Character Voice* than as *Narrator Voice* and are moderated according to the $ratio$ statistics, the *Character Voices* must take additional care to take into consideration appropriate character emotion or appraisal to be pragmatically cohesive.

$\mathbf{H}_3$ claims deaggregation and discourse structuring will effect reader preferences, and we observe a statistically significant difference in preference between stories generated by the narratological structurer's discourse structuring and baseline stories ((1, N=52) = 4.2, $p < 0.01$). There is no effect of story content or interaction. Table 25 shows examples of story segments generating with constructions that are preferred over the baseline. Deaggregation and discourse structuring were observed to create cleaner and crisper stories because they are shorter compared to the baseline stories, but still conveyed the point (e.g., the rank condition pair 2 in Table 25).

Pair	Rank Condition	Baseline Condition
1	Because the bugs scared John, he grabbed the rolled comic book.	John grabbed the rolled comic book be- cause the bugs scared him.
2	The squirrel fell over the deck's railing.	The squirrel fell over the deck's railing because the squirrel leaped because the squirrel was startled.
3	The squirrel was startled, so the squirrel leaped.	The squirrel leaped because the squirrel was startled.

Table 25: Excerpts of different discourse structuring conditions in rank vs. baseline ablation test

Rank vs. Random. Next, stories generated according to the $ratio$ statistics are compared against stories created with randomly assigned parameter values. For example, in a $ratio$ controlled story, an acknowledgement may only appear at most two times in *Character Voice*, whereas, in a random story, the same acknowledgement may appear in *Character Voice*, for example, four times. Ten story pairs of these types are annotated by the same seven Mechanical Turkers. Stories generated with $ratio$ statistics were preferred 92% of the time. Subjects found stories constructed using the statistical models easier to read and understand.

Random vs. Baseline. Finally, the stories with randomly assigned parameter values are compared against baseline stories with no sentence planning variation. Six story pairs of these types are annotated by the same seven Mechanical Turkers. The baseline stories were preferred 96% of the time. Subjects noted that no variation is better than poor variation, even if the baseline stories were plain. In one pair, the random story contained direct speech, a high ranked $ratio$ feature, with pragmatic markers in the direct speech. However, subjects found the dialogue to be stilted and did not contribute to a "good" character voice, and therefore preferred no variation. In the cases where a subject preferred the random story, it was stated that the baseline was "too dry".

Summary. These experiments show that the $catRatio$ and $perStoryRatio$ statistics learned from the overgeneration and rank experiments ensure consistency in the narratological structurer's generation throughout a story, and provide insight into how the narrative variations from the sentence planner can be combined to be most effective. These stories are preferable to randomly constructed stories. We also learn that no narrative variation is typically preferred to no poorly controlled random variation. In summary, the parameterized narratological structurer ensures that not only are generated stories unique, but that they will follow the standards of consistency and balance learned from observational data.

5.4 Predicting Selected Sentences

In addition to the statistical selection of the narratological structurer, we seek to determine whether we can learn a ranking, or preference function, that would allow us to select better sentences when generating stories, in a similar way to previous work (Walker et al., 2007). The findings here would be complementary to $catRatio$ and $perStoryRatio$, and used to make more informed decisions when generating during the Overgeneration phase.

We randomly split the 140 sentences used in the "Create Your Own Story" experiment into 100 for training and 40 for test. Each sentence was counted for how many times it was selected by annotators and binned as a binary Selected or Not-Selected class.[9] The task is classification: for each sentence, predict if it was selected by the annotators. We develop several feature sets with binary values to capture key aspects of these sentences:

- **N-gram features**: unigrams, bigrams, and trigrams.
- **Punct features**: punctuation marks.
- **Pragmatic features**: insertion of pragmatic markers, e.g., $ack : ok = True$.
- **Narrative features**: discourse structuring, point of view, and direct speech parameters.

An additional feature set, Pragmatic*, was created to further capture information about the context of these parameters. Rather than a binary "present" or "not present", Pragmatic* features represents the count of the features present. For example, the pragmatic feature $ack : ok = True$ for the Pragmatic feature set could be $ack : ok = 2$ for Pragmatic* feature set.

We train multiple off-the-shelf classification models using Weka.[10] The best models were Weka's Support Vector Machine implementation, SMO, and its decision tree, J48.[11] Tables 26 and 27 show the results on the training set using combinations of our features sets. The majority class baseline for predicting Selected (Sel.) is 0.4 and Not-Selected (¬Sel.) is 0.6. In both models, the "punct" features perform the best; for J48, no other features improve over the performance of "punct" alone, and for SMO, "punct" performs well in combination with "prag" and "ngram".

"N-gram" performs poorly on its own, which we posit could be due to the small size of the dataset or diversity of story content and domain specific vocabulary. We expected the narrative features to be useful because they represent more abstract narrative information, however on their

9. We note that the selected labels might be a result of the least-worst sentences, i.e., stories that were not generated with the best $catRatio$ and $perStoryRatio$ metrics, but this would still yield a subset of better potential sentences.
10. `https://www.cs.waikato.ac.nz/ml/weka/`
11. Other models tested include Weka's Naïve Bayes and Multilayer Perceptron.

Feature Set	Precision			Recall			F-Measure		
	Sel.	¬Sel.	Avg.	Sel.	¬Sel.	Avg.	Sel.	¬Sel.	Avg.
ngram	0.39	0.70	0.55	0.42	0.67	0.55	0.41	0.69	0.55
punct	0.48	0.72	0.60	0.36	0.81	0.59	0.41	0.76	0.59
prag	0.09	0.64	0.37	0.03	0.85	0.44	0.05	0.73	0.39
narr	0.00	0.67	0.34	0.00	1.00	0.50	0.00	0.80	0.40
ngram-narr	0.40	0.71	0.55	0.46	0.66	0.56	0.42	0.68	0.55
ngram-punct	0.45	0.74	0.59	0.52	0.69	0.60	0.48	0.71	0.60
ngram-prag	0.39	0.70	0.55	0.42	0.67	0.55	0.41	0.69	0.55
punct-narr	0.48	0.72	0.60	0.36	0.81	0.59	0.41	0.76	0.59
prag-narr	0.00	0.63	0.32	0.00	0.85	0.43	0.00	0.73	0.36
prag-punct	0.48	0.73	0.61	0.42	0.78	0.60	0.45	0.75	0.60
ngram-prag-narr	0.40	0.71	0.55	0.46	0.66	0.56	0.42	0.68	0.55
ngram-prag-puct	0.43	0.73	0.58	0.49	0.69	0.59	0.46	0.71	0.58
prag-punct-narr	0.52	0.75	0.63	0.46	0.79	0.62	0.48	0.77	**0.63**
ngram-prag-punct-narr	0.43	0.73	0.58	0.49	0.69	0.59	0.46	0.71	0.58

Table 26: Training Classification with SMO in Weka (highest averaged class f-measure in bold)

Feature Set	Precision			Recall			F-Measure		
	Sel.	¬Sel.	Avg.	Sel.	¬Sel.	Avg.	Sel.	¬Sel.	Avg.
ngram	0.27	0.66	0.46	0.12	0.84	0.48	0.17	0.74	0.45
punct	0.55	0.77	0.66	0.52	0.79	0.65	0.53	0.78	**0.66**
prag	0.14	0.66	0.40	0.03	0.91	0.47	0.05	0.76	0.41
narr	0.00	0.67	0.34	0.00	1.00	0.50	0.00	0.80	0.40
ngram-narr	0.27	0.66	0.46	0.12	0.84	0.48	0.17	0.74	0.45
ngram-punct	0.43	0.71	0.57	0.36	0.76	0.56	0.39	0.73	0.56
ngram-prag	0.27	0.66	0.46	0.12	0.84	0.48	0.17	0.74	0.45
punct-narr	0.52	0.75	0.63	0.46	0.79	0.62	0.48	0.77	0.63
prag-narr	0.00	0.66	0.33	0.00	0.96	0.48	0.00	0.78	0.39
prag-punct	0.53	0.78	0.65	0.58	0.75	0.66	0.55	0.76	**0.66**
ngram-prag-narr	0.27	0.66	0.46	0.12	0.84	0.48	0.17	0.74	0.45
ngram-prag-puct	0.43	0.71	0.57	0.36	0.76	0.56	0.39	0.73	0.56
prag-punct-narr	0.53	0.78	0.65	0.58	0.75	0.66	0.55	0.76	**0.66**
ngram-prag-punct-narr	0.43	0.71	0.57	0.36	0.76	0.56	0.39	0.73	0.56

Table 27: Training Classification with J48 in Weka (highest averaged class f-measure in bold)

own, the narrative features are not informative, and in some cases, bring down the scores when combined with other feature sets. We observe a similar phenomena in the interactions between realization and story content from our subjective experimentation: we cannot look at the realizations alone, but must take context and lexical realizations into consideration. We posit this is also why the Pragmatic features are not particularly informative either. The Pragmatic* features do not perform any better than the binary Pragmatic features and are excluded from the tables. When all features sets are combined together, they achieve a worse performance than the otherwise best performing feature sets, "prag-punct-narr" for SMO, and "punct" for the decision tree.

All models and feature sets are better at predicting the Not-Selected class (highest f-measure is 0.80 in Table 27), whereas Selected is more difficult to predict (highest f-measure is 0.55 in

Table 27). We posit that because the n-gram, pragmatic, and punct features are lexical, it is easier to categorize which lexicalizations are not preferred (e.g., the presence of the emphasizer *basically*, which had a $catRatio$ of 0, was never selected). However, of the lexical features that remain, it is more difficult to predict which would be selected. Even though these results seem high, the most informative features did not reveal general insights. For example, the most informative positive feature for the SMO classifier is "door" which is unique to a particular story in the training set, suggestive of overfitting. The decision tree reveals similar insights with a very deep and narrow tree.

Feature Set	Precision			Recall			F-Measure		
	Sel.	¬Sel.	Avg.	Sel.	¬Sel.	Avg.	Sel.	¬Sel.	Avg.
punct	0.64	0.86	0.75	0.64	0.86	0.75	0.64	0.86	0.75

Table 28: Testing Classification with J48 in Weka

We used the J48 prediction model with the best feature set, the simple "punct", on the test set of the remaining 40 sentences. The results are presented in Table 28 and show overall improvement to all the metrics. However, this best performing feature set only measures the appearance and type of punctuation generated in a sentence and leave much still to be understood about the features in this classification task. These results indicate that punctuation is key in determining which sentences will be selected or not, but this decision is the same function of the $perStoryRatio$ metric developed earlier, and thus this high-scoring feature set does not provide new insights. Another interpretation is that the features sets developed for this classification task were not able to truly capture what makes a sentence appealing enough to be Selected by an annotator, as seen by the wide the range of scores from the training metrics. A final interpretation is that there is simply not enough data or diversity for applying this type of model to this test set of Create-Your-Own-Stories for classification.

6. Conclusion and Future Work

This article has outlined requirements for a story planning and natural language generation story-telling system that bridges the four elements of the NLG story gap introduced in Section 2. We bridge the gap by designing the Fabula Tales to automatically map from a SIG (which can be obtained without domain knowledge) to a Lexical-Semantic representation (LSSTREE) compatible with a parameterized, narrative focused sentence planner. We train a narratological structurer from overgenerate and rank experimentation and observe trends from subject feedback. After the annotation of the SIG, the remainder of the translation and generation is streamlined.

Evaluation has shown that while the realizations produced with narrative variations are more effective than the baseline realizations, there is room for improvement with respect to fluency. Additional rules and heuristics can supplement the sentence planner to take context into consideration. Successful approaches have incorporated context with a rule-based approach, building on the senses in WordNet or VerbNet, as well as a statistical approach for expressive generation (Ahn et al., 2016; Rieser and Lemon, 2011; Paiva and Evans, 2004; Langkilde, 2000; Rowe et al., 2008; Mairesse and Walker, 2011). Fabula Tales currently does not examine domain or story specific knowledge during deaggregation or discourse structuring, but we posit that a closer examination of ontologies can be used to learn domain specific information from each story to influence its retelling. This would

allow, for example, the same content to be removed if it could be easily inferred from the prior context.

Fabula Tales requires manual annotation for the creation of SIGs, but we have shown that the *Scheherazade* tool is intuitive and lightweight. While some stories are more difficult than others to annotate, the guidelines we have adopted show strategies for encoding these interpretations. Utilizing SIGs affords manipulation of many aspects of the narrative that we have not yet explored, including *Time* variations beyond the straightforward temporal ordering. Bae et al. (2011) implement a computational model of focalization in generating narratives, where a planning-based generation engine identifies which events or inner thoughts characters are aware at the moment. These in turn, affect the selection of the events the planner selects to tell. For example, stories with surprise endings are generated by exploiting the disparity of knowledge between a story's reader and its characters (Bae and Young, 2009). Other work describes narrative systems that avoid conflict or make assumptions about its structure, or rely on humans to author it, creating a system based on character worlds, plans, their intentionality, and goals (Ware and Young, 2011; Ware et al., 2014; Ware and Young, 2012). Characters' plans, intentions, and inner thoughts are annotated in the INTERPRETATION and AFFECTUAL layers of the SIG, which we posit can be used to explore these narrative aspects. Furthermore, we have assumed that a single SIG represents the whole of the *fabula*. Future work may examine overlap of SIGs derived from the same story, such as the collection of fables that have multiple SIGs in the DramaBank, and explore how to combine information from different SIGs and create a model for focalization.

There are several promising lines of work that can be explored with our ability to now bridge the NLG story gap for integrated applications. Personalization can lead to even more engagement, and especially by coordinating gestures with speech of embodied virtual agents to increase the naturalness of human-like communication (Hu et al., 2015; Bergmann et al., 2013; Wang and Neff, 2013). The Fabula Tales framework could be integrated into a virtual agent environment to provide a plethora of stories for the agent to tell; as such, the affordances of the sentence-level variations is taken advantage of by rendering dialogue between two virtual agents (Hu et al., 2016). Here also, the gesture generation would benefit from the integration of story-level interactions. This NLG storytelling framework could also be used to enhance narrative systems, as well as make them more customized to the user, especially in simulated scenarios (Johnson et al., 2004; Aylett et al., 2005). Stories written by real people could be generated from different perspectives to further explore perceptions and empathy by changing point of view parameter and pragmatic markers.

Another application area is narratively structured computer games. Dialogue authoring in large games requires not only the creation of new content, but the subtlety of its delivery as it varies from character to character. When creating a replayable and adaptable system, it is important to have believable interactions with non-player characters (NPCs) in the game world. Short of hand authoring every possible character utterance, we ask "can NPCs be given a personality fitted to the player?" and "upon replay, do the NPCs utterances change under the assumption that the story is the same?" Expanding NPC dialogue generation was explored in Lukin et al. (2014), but extending this work requires a fully immersive game world, characters, and improvement to Fabula Tales' story-level planner in order to show its effectiveness. Along these lines, dialogue variants corresponding to in-game regional dialects could be modeled. Some work has made great strides toward richer modeling of social-group membership for virtual characters (Harrell et al., 2014; Walker et al., 2013), and the ability to automatically produce linguistic variation according to such models would greatly enhance the impact of the systems.

7. Acknowledgments

This work was partially funded by an ARCS Foundation scholarship, a Nuance Communications fellowship, and a University of California, Santa Cruz Baskin Family fellowship. Research was partially sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-17-2-0064. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. We thank the anonymous reviewers for their constructive and helpful feedback.

Appendix A. Statistical Tests

Significance tests for Sections 4.4.1 and 4.4.2.

Style	1. Original	2. 1st-out	3. 1st-neutr	4. 1st-shy	5. *Scheherazade*
1. Original	–				
2. 1st-out	-4.00***	–			
3. 1st-neutr	-7.09****	-1.63	–		
4. 1st-shy	-8.02****	8.93****	3.72***	–	
5. *Scheherazade*	14.35****	5.61****	7.32****	6.03***	–
6. 3rd-neutr	-13.97****	9.50****	8.30****	3.33***	-0.31

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$

Table 29: Point of View and Voice t-values for Engagement (df = 95)

Style	1. Original	2. 1st-out	3. 1st-neutr	4. 1st-shy	5. *Scheherazade*
1. Original	–				
2. 1st-out	5.59****	–			
3. 1st-neutr	6.05****	> 1	–		
4. 1st-shy	6.90****	1.51	2.20*	–	
5. *Scheherazade*	14.46****	7.28****	7.58****	6.16****	–
6. 3rd-neutr	14.24****	7.75****	8.34****	6.89****	0.54

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$

Table 30: Point of View and Voice t-values for Interest (df = 93)

Realization	1. Original	2. soSN	3 becauseNS	4 becauseSN	5. NS	6. None
1. Original	–					
2. soSN	2.6**	–				
3. becauseNS	3.2***	-1.1	–			
4. becauseSN	3.1**	-1.3*	0.1*	–		
5. NS	4.3****	-3.0**	-2.2*	-2.2*	–	
6. None	4.9****	-3.0**	-2.1**	-2.4 **	-0.2	–
7. N	7.1****	-3.7***	-3.1***	-3.1 **	1.6	1.5

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$

Table 31: Deaggregation and Discourse Structuring t-values for Correctness, exp. 1 (df = 101)

Realization	1. Original	2. soSN	3 becauseNS	4 becauseSN	5. NS	6. None
1. Original	–					
2. soSN	-4.1****	–				
3. becauseNS	-5.4****	-2.4**	–			
4. becauseSN	-5.6****	-2.7**	-0.5	–		
5. NS	-7.6****	5.6****	3.3***	3.1**	–	
6. None	-9.4****	7.5****	5.7****	5.4****	2.7**	–
7. N	-10.8****	-6.1****	-4.3****	-3.8***	-1.7*	0.3

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$

Table 32: Deaggregation and Discourse Structuring t-values for Preference, exp. 1 (df = 106)

Realization	1. Original	2. soSN
1. Original	–	
2. soSN	< 1****	–
3. *Scheherazade*	< 1****	< 1****

**** $p < 0.0001$

Table 33: Deaggregation and Discourse Structuring t-values for Correctness, exp. 2 (df = 95)

Realization	1. Original	2. soSN
1. Original	–	
2. soSN	< 1****	–
3. *Scheherazade*	< 1****	< 1****

**** $p < 0.0001$

Table 34: Deaggregation and Discourse Structuring t-values for Preference, exp. 2 (df = 95)

References

H. Porter Abbott. *The Cambridge Introduction to Narrative*. Cambridge University Press, 2008.

Emily Ahn, Fabrizio Morbini, and Andrew S Gordon. Improving Fluency in Narrative Text Generation With Grammatical Transformations and Probabilistic Parsing. In *In Proc. of the 9th International Natural Language Generation conference*, page 70, 2016.

Ruth S Aylett, Sandy Louchart, Joao Dias, Ana Paiva, and Marco Vala. FearNot!–An Experiment in Emergent Narrative. In *Proc. of Intelligent Virtual Agents*, pages 305–316. Springer, 2005.

Byung-Chull Bae and R Michael Young. Suspense? Surprise! or How to Generate Stories with Surprise Endings by Exploiting the Disparity of Knowledge between a Story's Reader and its Characters. In *Proc. of Interactive Storytelling*, pages 304–307. Springer Berlin Heidelberg, 2009.

Byung-Chull Bae, Yun-Gyung Cheong, and R. Michael Young. Toward a Computational Model of Focalization in Narrative. In *Proc. of the 6th International Conference on Foundations of Digital Games*, pages 313–315. ACM, 2011.

Mieke Bal. Narratology. Introduction to the Theory of Narrative. 1997.

Camille Barot, Colin M Potts, and R Michael Young. A Tripartite Plan-Based Model of Narrative for Narrative Discourse Generation. In *Proc. of the Joint Workshop on Intelligent Narrative Technologies and Social Believability in Games at the 11th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, pages 2–8, 2015.

Anja Belz and Ehud Reiter. Comparing Automatic and Human Evaluation of NLG Systems. In *Proc. of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006.

Kirsten Bergmann, Sebastian Kahl, and Stefan Kopp. Modeling the Semantic Coordination of Speech and Gesture under Cognitive and Linguistic Constraints. In *Proc. of Intelligent Virtual Agents*, pages 203–216. Springer, 2013.

Douglas Biber. *Variation Across Speech and Writing*. Cambridge University Press, 1991.

Nadjet Bouayad-Agha, Donia R Scott, and Richard Power. Integrating Content and Style in Documents: A Case Study of Patient Information Leaflets. *Information Design Journal*, 9(2-3): 161–176, 1998.

Jerome Bruner. The Narrative Construction of Reality. *Critical Inquiry*, 18:1–21, 1991.

Kevin Burton, Akshay Java, Ian Soboroff, et al. The ICWSM 2009 Spinn3r Dataset. In *Proc. of the Third Annual Conference on Weblogs and Social Media*, 2009.

Lynne Cahill, John Carroll, Roger Evans, Daniel Paiva, Richard Power, Donia Scott, and Kees van Deemter. From RAGS to RICHES: Exploiting the Potential of a Flexible Generation Architecture. In *Proc. of the 39th Annual Meeting on Association for Computational Linguistics*, pages 106–113. Association for Computational Linguistics, 2001.

Charles B Callaway and James C Lester. Narrative Prose Generation. *Artificial Intelligence*, 139 (2):213–252, 2002.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 Task 1: Semantic Textual Similarity-Multilingual and Cross-lingual Focused Evaluation. 2017.

Raman Chandrasekar and Bangalore Srinivas. Automatic Induction of Rules for Text Simplification. *Knowledge-Based Systems*, 10(3):183–190, 1997.

Yun-Gyung Cheong and R Michael Young. Narrative Generation for Suspense: Modeling and Evaluation. In *Proc. of the International Conference on Interactive Digital Storytelling*, 2008.

Eugenio Concepción, Pablo Gervás, Gonzalo Méndez, and Carlos León. Using CNL for Knowledge Elicitation and Exchange across Story Generation Systems. In *Proc. of the International Workshop on Controlled Natural Language*, pages 81–91. Springer, 2016a.

Eugenio Concepción, Gonzalo Mendez, and Pablo Gervás. Mining Knowledge in Storytelling Systems for Narrative Generation. In *Proc. of the Workshop on Computational Creativity in Natural Language Generation at the International Natual Langauge Generation Conference*, pages 41–50, 2016b.

Eugenio Concepción, Gonzalo Méndez, Pablo Gervás, and Carlos León. A Challenge Proposal for Narrative Generation Using CNLs. In *Proc. of the 9th International Natural Language Generation Conference*, pages 171–173, 2016c.

Michael Denkowski and Alon Lavie. METEOR Universal: Language Specific Translation Evaluation for Any Target Language. In *Proc. of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, 2014.

George Doddington. Automatic Evaluation of Machine Translation Quality using N-gram Co-occurrence Statistics. In *Proc. of the Second International Conference on Human Language Technology Research*, pages 138–145. Morgan Kaufmann Publishers Inc., 2002.

Michael Elhadad and Jacques Robin. An Overview of SURGE: A Reusable Comprehensive Syntactic Realization Component. Technical report, Technical Report 96-03, Ben Gurion University, Dept. of Computer Science, Beer Sheva, Israel, 1996.

David Elson. *Modeling Narrative Discourse*. PhD thesis, Columbia University, Dept. of Computer Science, 2012a.

David K Elson. DramaBank: Annotating Agency in Narrative Discourse. In *Proc. of the 8th International Conference on Language Resources and Evaluation*, 2012b.

David K Elson and Kathleen R McKeown. A Tool for Deep Semantic Encoding of Narrative Texts. In *Proc. of the ACL-IJCNLP 2009 Software Demonstrations*, pages 9–12. Association for Computational Linguistics, 2009.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical Neural Story Generation. In *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 889–898. Association for Computational Linguistics, 2018.

Christiane Fellbaum. WordNet: An Electronic Lexical Database. *WordNet is available from http://www. cogsci. princeton. edu/wn*, 2010.

Katja Filippova and Michael Strube. Sentence Fusion via Dependency Graph Compression. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pages 177–185. Association for Computational Linguistics, 2008.

Claire Gardent and German Kruszewski. Generation for Grammar Engineering. In *Proc. of the Seventh International Natural Language Generation Conference*, pages 31–39. Association for Computational Linguistics, 2012.

Gérard Genette and Jane E Lewin. *Narrative Discourse: An Essay in Method*. Cornell University Press, 1983.

Richard J Gerrig. *Experiencing Narrative Worlds: On the Psychological Activities of Reading*. Yale University Press, 1993.

Pablo Gervás, Belén Díaz-Agudo, Federico Peinado, and Raquel Hervás. Story Plot Generation Based on CBR. In *In Proc. of Applications and Innovations in Intelligent Systems XII*, pages 33–46. Springer, 2005.

Pablo Gervás, Birte Lönneker-Rodman, Jan Christoph Meister, and Federico Peinado. Narrative Models: Narratology Meets Artificial Intelligence. In *Proc. of the Toward Computational Models of Literary Analysis Work at the International Conference on Language Resources and Evaluation*, pages 44–51, 2006.

Gregory Grefenstette. Producing Intelligent Telegraphic Text Reduction to Provide an Audio Scanning Service for the Blind. In *working notes of the AAAI Spring Symposium on Intelligent Text summarization*, pages 111–118, 1998.

D Fox Harrell, Dominic Kao, Chong-U Lim, Jason Lipshin, Ainsley Sutherland, and Julia Makivic. The Chimeria Platform: An Intelligent Narrative System for Modeling Social Identity-Related Experiences. In *Proc. of the Seventh Intelligent Narrative Technologies Workshop*, 2014.

David Howcroft, Crystal Nakatsu, and Michael White. Enhancing the Expression of Contrast in the SPaRKy Restaurant Corpus. In *Proc. of the 14th European Workshop on Natural Language Generation*, pages 30–39, 2013.

Chao Hu, Marilyn A Walker, Michael Neff, and Jean E Fox Tree. Storytelling Agents with Personality and Adaptivity. In *Proc. of the International Conference on Intelligent Virtual Agents*, pages 181–193. Springer, 2015.

Zhichao Hu, Elahe Rahimtoroghi, Larissa Munishkina, Reid Swanson, and Marilyn A Walker. Unsupervised Induction of Contingent Event Pairs from Film Scenes. In *Proc. of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 369–379, 2013.

Zhichao Hu, Michelle Dick, Chung-Ning Chang, Kevin Bowden, Michael Neff, Jean E Fox Tree, and Marilyn A Walker. A Corpus of Gesture-Annotated Dialogs for Monologue-to-Dialogue Generation from Personal Narratives. In *Proc. of the 10th International Conference on Language Resources and Evaluation*, 2016.

Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual Storytelling. In *Proc. of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239, 2016.

Diana Zaiu Inkpen and Graeme Hirst. Near-Synonym Choice in Natural Language Generation. In *Recent Advances in Natural Language Processing*, volume 3, pages 141–152, 2004.

Amy Isard, Carsten Brockmann, and Jon Oberlander. Individuality and Alignment in Generated Dialogues. *Proc. of the 4th International Natural Language Generation Conference*, 2006.

W Lewis Johnson, Carole Beal, Anna Fowles-Winkler, Ursula Lauper, Stacy Marsella, Shrikanth Narayanan, Dimitra Papachristou, and Hannes Vilhjálmsson. Tactical Language Training System: An Interim Report. In *Proc. of Intelligent Tutoring Systems*, pages 336–345. Springer, 2004.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. Extensive Classifications of English Verbs. In *Proc. of the 12th EURALEX International Congress*, pages 1–15, 2006.

Kevin Knight and Daniel Marcu. Statistics-based summarization-step one: Sentence compression. *Proc. of AAAI/IAAI*, 2000:703–710, 2000.

Kevin Knight and Daniel Marcu. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence*, 139(1):91–107, 2002.

William Labov and Joshua Waletzky. Narrative analysis: Oral Versions of Personal Experience. 1997.

Irene Langkilde. Forest-Based Statistical Sentence Generation. In *Proc. of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 170–177. Association for Computational Linguistics, 2000.

Irene Langkilde and Kevin Knight. Generation that Exploits Corpus-Based Statistical Knowledge. In *Proc. of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, pages 704–710. Association for Computational Linguistics, 1998.

Irene Langkilde-Geary. An Empirical Verification of Coverage and Correctness for a General-Purpose Sentence Generator. In *Proc. of the 12th International Natural Language Generation Workshop*, pages 17–24, 2002.

Benoit Lavoie and Owen Rambow. A Fast and Portable Realizer for Text Generation Systems. In *Proc. of the fifth conference on Applied natural language processing*, pages 265–268. Association for Computational Linguistics, 1997.

Michael Lebowitz. Creating a Story-Telling Universe. In *Proc. of International Joint Conferences on Artificial Intelligence*, pages 63–65. Citeseer, 1983.

Boyang Li. *Learning Knowledge to Support Domain-Independent Narrative Intelligence*. PhD thesis, Georgia Institute of Technology, 2015.

Boyang Li, Stephen Lee-Urban, George Johnston, and Mark O Riedl. Story Generation with Crowdsourced Plot Graphs. In *Proc. of the 27th AAAI Conferece on Artificial Intelligence*, 2013.

Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. 2004.

Grace Lin. *Character Modeling through Dialogue for Expressive Natural Language Generation*. PhD thesis, University of California, Santa Cruz, Dept. of Computer Science, 2016.

Birte Lönneker. Narratological Knowledge for Natural Language Generation. In *Proc. of the 10th European Workshop on Natural Language Generation*, pages 91–100. Citeseer, 2005.

Stephanie Lukin, Reginald Hobbs, and Clare Voss. A Pipeline for Creative Visual Storytelling. In *Proc. of the First Workshop on Storytelling*, pages 20–32, 2018.

Stephanie M Lukin and Marilyn A Walker. Narrative Variations in a Virtual Storyteller. In *Proc. of Intelligent Virtual Agents*, pages 320–331. Springer, 2015.

Stephanie M Lukin, James O Ryan, and Marilyn A Walker. Automating Direct Speech Variations in Stories and Games. In *Proc. of the Workshop on Games and NLP at the Tenth Artificial Intelligence and Interactive Digital Entertainment Conference*, 2014.

Stephanie M Lukin, Lena I Reed, and Marilyn A Walker. Generating Sentence Planning Variations for Story Telling. In *Proc of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 188, 2015.

Stephanie M. Lukin, Kevin Bowden, Casey Barackman, and Marilyn A. Walker. PersonaBank: A Corpus of Personal Narratives and Their Story Intention Graphs. In *Proc. of the 10th International Conference on Language Resources and Evaluation*, 2016.

Matt Madden. *99 Ways to Tell a Story*. Random House, 2006.

F. Mairesse and M.A. Walker. Towards personality-based user adaptation: psychologically informed stylistic language generation. *User Modeling and User-Adapted Interaction*, pages 1–52, 2010a. ISSN 0924-1868.

François Mairesse and Marilyn A Walker. A Personality-Based Framework for Utterance Generation in Dialogue Applications. In *AAAI Spring Symposium: Emotion, Personality, and Social Behavior*, pages 80–87, 2008.

François Mairesse and Marilyn A Walker. Towards Personality-Based User Adaptation: Psychologically Informed Stylistic Language Generation. *User Modeling and User-Adapted Interaction*, 20(3):227–278, 2010b.

François Mairesse and Marilyn A Walker. Controlling User Perceptions of Linguistic Style: Trainable Generation of Personality Traits. *Computational Linguistics*, 37(3):455–488, 2011.

William C Mann and Sandra A Thompson. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3): 243–281, 1988.

Erwin Marsi and Emiel Krahmer. Explorations in sentence fusion. In *Proc. of the Tenth European Workshop on Natural Language Generation (ENLG-05)*, 2005.

Michael Mateas. A Preliminary Poetics for Interactive Drama and Games. *Digital Creativity*, 12 (3):140–152, 2001.

Dan P McAdams, Ruthellen Ed Josselson, and Amia Ed Lieblich. *Identity and Story: Creating Self in Narrative.* American Psychological Association, 2006.

James R Meehan. TALE-SPIN, An Interactive Program that Writes Stories. In *Proc. of the International Joint Conferences on Artificial Intelligence*, volume 77, pages 91–98. Citeseer, 1977.

Igor A. Mel'čuk. *Dependency Syntax: Theory and Practice*. SUNY Press, 1988.

Nick Montfort. *Generating Narrative Variation in Interactive Fiction*. PhD thesis, University of Pennsylvania, Dept. of Computer and Information Science, 2007.

Nick Montfort. Curveship: an Interactive Fiction System for Interactive Narrating. In *Proc. of the Workshop on Computational Approaches to Linguistic Creativity*, pages 55–62. Association for Computational Linguistics, 2009.

Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. LSDSem 2017 Shared Task: The Story Cloze Test. In *Proc. of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 46–51, 2017.

Larissa Munishkina, Jennifer Parrish, and Marilyn A Walker. Fully-Automatic Interactive Story Design from Film Scripts. In *Proc. of the International Conference on Interactive Digital Storytelling*, pages 229–232. Springer, 2013.

Shashi Narayan, Claire Gardent, Shay Cohen, and Anastasia Shimorina. Split and Rephrase. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pages 617–627, 2017.

James Niehaus and R Michael Young. A Computational Model of Inferencing in Narrative. In *AAAI Spring Symposium: Intelligent Narrative Technologies II*, pages 83–90, 2009.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. Why We Need New Evaluation Metrics for NLG. In *Proc. of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, 2017.

Shereen Oraby, Lena Reed, Shubhangi Tandon, TS Sharath, Stephanie Lukin, and Marilyn Walker. Controlling Personality-Based Stylistic Variation with Neural Natural Language Generators. In *Proc. of the Special Interest Group on Discourse and Dialogue*, 2018.

Daniel S. Paiva and Roger Evans. A framework for stylistically controlled generation. In Anja Belz, Roger Evans, and Paul Piwek, editors, *Natural Language Generation, Third Internatonal Conference, INLG 2004*, number 3123 in LNAI, pages 120–129. Springer, July 2004.

Alan Palmer. Universal Minds. *Semiotica*, 2007(165):205–225, 2007.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proc. of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

Cécile Paris and Donia Scott. Stylistic Variation in Multilingual Instructions. In *Proc. of the Seventh International Workshop on Natural Language Generation*, pages 45–52. Association for Computational Linguistics, 1994.

Federico Peinado and Pablo Gervás. Evaluation of Automatic Generation of Basic Stories. *New Generation Computing*, 24(3):289–302, 2006.

Nanyun Peng, Marjan Ghazvininejad, Jonathan May, and Kevin Knight. Towards Controllable Story Generation. In *Proc. of the First Workshop on Storytelling*, pages 43–49, 2018.

Manon Penning and Mariët Theune. Cueing the Virtual Storyteller: Analysis of cue phrase usage in fairy tales. In *Proc. of the Eleventh European Workshop on Natural Language Generation*, pages 159–162. Association for Computational Linguistics, 2007.

David Pizarro, Eric Uhlmann, and Peter Salovey. Asymmetry in Judgments of Moral Blame and Praise The Role of Perceived Metadesires. *Psychological Science*, 14(3):267–272, 2003.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The Penn Discourse TreeBank 2.0. In *Proc. of 6th International Conference on Language Resources and Evaluation*, 2008.

Gerald Prince. *A Grammar of Stories: An Introduction*, volume 13. Walter de Gruyter, 1974.

Vladimir Propp. *Morphology of the Folktale*. University of Texas Press, second edition, 1969.

Raymond Queneau and Barbara Wright. *Exercises in Style*, volume 513. New Directions Publishing, 1981.

Elahe Rahimtoroghi, Reid Swanson, Marilyn A Walker, and Thomas Corcoran. Evaluation, Orientation, and Action in Interactive Storytelling. In *Proc. of Intelligent Narrative Technologies Workshop*, volume 6, 2013.

Elahe Rahimtoroghi, Thomas Corcoran, Reid Swanson, Marilyn A Walker, Kenji Sagae, and Andrew Gordon. Minimal Narrative Annotation Schemes and their Applications. In *Proc. of the Intelligent Narrative Technologies Workshop*, 2014.

Elahe Rahimtoroghi, Ernesto Hernandez, and Marilyn Walker. Learning Fine-Grained Knowledge about Contingent Relations between Everyday Events. In *Proc. of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 350–359, 2016.

Aaron A Reed, Ben Samuel, Anne Sullivan, Ricky Grant, April Grow, Justin Lazaro, Jennifer Mahal, Sri Kurniawan, Marilyn A Walker, and Noah Wardrip-Fruin. A Step Towards the Future of Role-Playing Games: The SpyFeet Mobile RPG Project. In *Proc. of the Conference on Artificial Intelligence and Interactive Digital Entertainment Conference*, 2011.

Mark O Riedl and Robert Michael Young. Narrative Planning: Balancing Plot and Character. *Journal of Artificial Intelligence Research*, 39(1):217–268, 2010.

Mark Owen Riedl and R Michael Young. An Intent-Driven Planner for Multi-Agent Story Generation. In *Proc. of the Third International Joint Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 186–193. IEEE Computer Society, 2004.

Verena Rieser and Oliver Lemon. *Reinforcement Learning for Adaptive Dialogue Systems: A Data-Driven Methodology for Dialogue Management and Natural Language Generation*. Springer Science & Business Media, 2011.

Stefan Riezler, Tracy H King, Richard Crouch, and Annie Zaenen. Statistical sentence condensation using ambiguity packing and stochastic disambiguation methods for lexical-functional grammar. In *Proc. of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 118–125. Association for Computational Linguistics, 2003.

Elena Rishes, Stephanie M Lukin, David K Elson, and Marilyn A Walker. Generating Different Story Tellings from Semantic Representations of Narrative. In *Proc. of Interactive Storytelling*, pages 192–204. Springer International Publishing, 2013.

Melissa Roemmele. *Neural Networks for Narrative Continuation*. PhD thesis, Ph. D. Dissertation, University of Southern California, 2018.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, pages 90–95, 2011.

Jonathan P Rowe, Eun Young Ha, and James C Lester. Archetype-Driven Character Dialogue Generation for Interactive Narrative. In *Proc. of Intelligent Virtual Agents*, pages 45–58. Springer, 2008.

Donia Scott and Clarisse Sieckenius de Souza. Getting the Message Across in RST-Based Text Generation. *Current Research in Natural Language Generation*, 4:47–73, 1990.

Rishi Sharma, James Allen, Omid Bakhshandeh, and Nasrin Mostafazadeh. Tackling the Story Ending Biases in The Story Cloze Test. In *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 752–757. Association for Computational Linguistics, 2018.

Siddarth Srinivasan, Richa Arora, and Mark Riedl. A Simple and Effective Approach to the Story Cloze Test. In *Proc. of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 92–96. Association for Computational Linguistics, 2018.

Reid Swanson and Andrew S Gordon. Say Anything: A Massively Collaborative Open Domain Story Writing Companion. In *Proc. of Interactive Storytelling*, pages 32–40. Springer, 2008.

Mariët Theune, Sander Rensen, Rieks op den Akker, Dirk Heylen, and Anton Nijholt. Emotional Characters for Automatic Plot Creation. In *Proc. of the International Conference on Technologies for Interactive Digital Storytelling and Entertainment*, pages 95–100. Springer, 2004.

Mariët Theune, Nanda Slabbers, and Feikje Hielkema. The Narrator: NLG for Digital Storytelling. In *Proc. of the Eleventh European Workshop on Natural Language Generation*, pages 109–112. Association for Computational Linguistics, 2007.

Avril Thorne. The Press of Personality: A Study of Conversations between Introverts and Extraverts. *Journal of Personality and Social Psychology*, 53(4):718, 1987.

Avril Thorne and Kate C McLean. Telling Traumatic Events in Adolescence: A Study of Master Narrative Positioning. *Connecting Culture and Memory: The Development of an Autobiographical Self*, pages 169–185, 2003.

Scott R Turner. Minstrel: A Computer Model of Creativity and Storytelling. 1993.

Marilyn Walker, Owen Rambow, and Monica Rogati. Training a Sentence Planner for Spoken Dialogue Using Boosting. *Computer Speech and Language: Special Issue on Spoken Language Generation*, 16(3-4):409–433, 2002.

Marilyn A Walker, Amanda Stent, François Mairesse, and Rashmi Prasad. Individual and Domain Adaptation in Sentence Planning for Dialogue. *Journal of Artificial Intelligence Research*, 30: 413–456, 2007.

Marilyn A Walker, Jennifer Sawyer, Carolynn Jimenez, Elena Rishes, Grace I Lin, Zhichao Hu, Jane Pinckard, and Noah Wardrip-Fruin. Using Expressive Language Generation to Increase Authorial Leverage. In *Proc. of the Ninth Artificial Intelligence and Interactive Digital Entertainment Conference*, 2013.

Xin Wang, Wenhu Chen, Yuan-Fang Wang, and William Yang Wang. No Metrics Are Perfect: Adversarial Reward Learning for Visual Storytelling. In *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 899–909. Association for Computational Linguistics, 2018.

Yingying Wang and Michael Neff. The Influence of Prosody on the Requirements for Gesture-Text Alignment. In *Proc. of Intelligent Virtual Agents*, pages 180–188. Springer, 2013.

Stephen G Ware and R Michael Young. Validating a Plan-Based Model of Narrative Conflict. In *Proc. of the International Conference on the Foundations of Digital Games*, pages 220–227. ACM, 2012.

Stephen G. Ware and Robert Michael Young. CPOCL: A Narrative Planner Supporting Conflict. In *Proc. of the Artificial Intelligence and Interactive Digital Entertainment Conference*, 2011.

Stephen G Ware, R Michael Young, Brent Harrison, and David L Roberts. A Computational Model of Plan-Based Narrative Conflict at the Fabula Level. *IEEE Transactions on Computational Intelligence and AI in Games*, 6(3):271–288, 2014.

David Winer and R Michael Young. Discourse-Driven Narrative Generation with Bipartite Planning. In *Proc. of the 9th International Natural Language Generation conference*, pages 11–20, 2016.