Generating Sentence Planning Variations for Story Telling

Stephanie M. Lukin, Lena I. Reed & Marilyn A. Walker
Natural Language and Dialogue Systems
University of California, Santa Cruz
Baskin School of Engineering
slukin,lireed,mawalker@ucsc.edu

Abstract

There has been a recent explosion in applications for dialogue interaction ranging from direction-giving and tourist information to interactive story systems. Yet the natural language generation (NLG) component for many of these systems remains largely handcrafted. This limitation greatly restricts the range of applications; it also means that it is impossible to take advantage of recent work in expressive and statistical language generation that can dynamically and automatically produce a large number of variations of given content. We propose that a solution to this problem lies in new methods for developing language generation resources. We describe the ES-TRANSLATOR, a computational language generator that has previously been applied only to fables, and quantitatively evaluate the domain independence of the EST by applying it to personal narratives from weblogs. We then take advantage of recent work on language generation to create a parameterized sentence planner for story generation that provides aggregation operations, variations in discourse and in point of view. Finally, we present a user evaluation of different personal narrative retellings.

1 Introduction

Recently there has been an explosion in applications for natural language and dialogue interaction ranging from direction-giving and tourist information to interactive story systems (Dethlefs et al., 2014; Walker et al., 2011; Hu et al., 2015). While this is due in part to progress in statistical natural language understanding, many applications require the system to actually respond in a meaningful way. Yet the natural language generation (NLG) component of many interactive dialogue systems remains largely handcrafted. This

Original
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 outside. His bowl has become a very popular site. Throughout the day, many birds drink out of it and bathe in it. The birds literally line up on the railing and wait their turn. Squirrels also come to drink out of it. 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!

Table 1: The Startled Squirrel Weblog Story

limitation greatly restricts the range of applications; it also means that it is impossible to take advantage of recent work in expressive and statistical language generation that can dynamically and automatically produce a large number of variations of given content (Rieser and Lemon, 2011; Paiva and Evans, 2004; Langkilde, 1998; Rowe et al., 2008; Mairesse and Walker, 2011). Such variations are important for expressive purposes, we well as for user adaptation and personalization (Zukerman and Litman, 2001; Wang et al., 2005; McQuiggan et al., 2008). We propose that a solution to this problem lies in new methods for developing language generation resources.

First we describe the ES-TRANSLATOR (or EST), a computational language generator that has previously been applied only to fables, e.g. the fable in Table 3 (Rishes et al., 2013). We quantitatively evaluate the domain independence of the EST by applying it to social media narratives, such as the *Startled Squirrel* story in Table 1. We then present a parameterized general-purpose framework built on the EST pipeline, EST 2.0, that can generate many different tellings of the same story, by utilizing sentence planning and point of view parameters. Automatically generated story variations are shown in Table 2 and Table 4.

We hypothesize many potential uses for our ap-

188

EST 2.0
Benjamin wanted to drink the bowl's water, so I placed the bowl on the deck. The bowl was popular. The birds drank the bowl's water. The birds bathed themselves in the bowl. The birds organized themselves on the deck's railing because the birds wanted to wait. The squirrels drank the bowl's water. The squirrel approached the bowl. The squirrel was startled because the squirrel saw the squirrel's reflection. Because it was startled, the squirrel leapt. 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 2: Retelling of the Startled Squirrel

Original
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 3: "The Fox and the Crow"

proach to repurposing and retelling existing stories. First, such stories are created daily in the thousands and cover any topic imaginable. They are natural and personal, and may be funny, sad, heart-warming or serious. There are many potential applications: virtual companions, educational storytelling, or to share troubles in therapeutic settings (Bickmore, 2003; Pennebaker and Seagal, 1999; Gratch et al., 2012).

Previous research on NLG of linguistic style shows that dialogue systems are more effective if they can generate stylistic linguistic variations based on the user's emotional state, personality, style, confidence, or other factors (André et al., 2000; Piwek, 2003; McQuiggan et al., 2008; Porayska-Pomsta and Mellish, 2004; Forbes-Riley and Litman, 2011; Wang et al., 2005; Dethlefs et al., 2014). Other work focuses on variation in journalistic writing or instruction manuals, where stylistic variations as well as journalistic slant or connotations have been explored (Hovy, 1988; Green and DiMarco, 1993; Paris and Scott, 1994; Power et al., 2003; Inkpen and Hirst, 2004). Previous iterations of the EST simply presented a sequence of events (Rishes et al., 2013). This work implements parameterized variation of linguistic style in the context of weblogs in order to introduce discourse structure into our generated stories.

Our approach differs from previous work on NLG for narrative because we emphasize (1) domain-independent methods; and (2) generating a large range of variation, both narratological and stylistic. (Lukin and Walker, 2015)'s work on the EST is the first to generate dialogue within stories, to have the ability to vary direct vs. indirect speech, and to generate dialogue utterances using different stylistic models for character voices. Previous work can generate narratological variations, but is domain dependent (Callaway and Lester, 2002; Montfort, 2007).

Sec. 2 describes our corpus of stories and the ar-

chitecture of our story generation framework, EST 2.0.[1] Sec. 3 describes experiments testing the coverage and correctness of EST 2.0. Sec. 4 describes experiments testing user perceptions of different linguistic variations in storytelling. Our contributions are:

- We produce SIG representations of 100 personal narratives from a weblog corpus, using the story annotation tool Scheherezade (Elson and McKeown, 2009; Elson, 2012);
- We compare EST 2.0 to EST and show how we have not only made improvements to the translation algorithm, but can extend and compare to personal narratives.
- We implement a parameterized variation of linguistic style in order to introduce discourse structure into our generated narratives.
- We carry out experiments to gather user perceptions of different sentence planning choices that can be made with complex sentences in stories.

We sum up and discuss future work in Sec. 5.
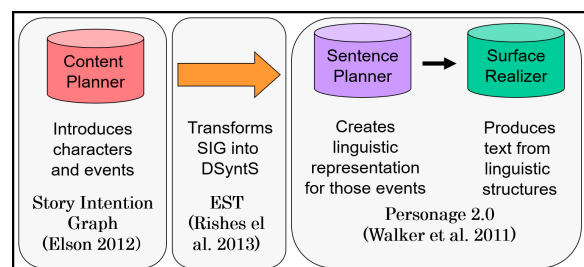
2 Story Generation Framework

Figure 1: NLG pipeline method of the ES Translator.

Fig. 1 illustrates our overall architecture, which uses NLG modules to separate the process of planning *What to say* (content planning and selection,

---

fabula) from decisions about *How to say it* (sentence planning and realization, discourse). We build on three existing tools from previous work: the SCHEHEREZADE story annotation tool, the PERSONAGE generator, and the ES-TRANSLATOR (EST) (Elson, 2012; Mairesse and Walker, 2011; Rishes et al., 2013). The EST uses the STORY INTENTION GRAPH (SIG) representation produced by SCHEHEREZADE and its theoretical grounding as a basis for the content for generation. The EST bridges the narrative representation of the SIG to the representation required by PERSONAGE by generating the text plans and the deep syntactic structures that PERSONAGE requires. Thus any story or content represented as a SIG can be retold using PERSONAGE. See Fig. 1.

There are several advantages to using the SIG as the representation for a content pool:

- Elson's DRAMABANK provides stories encoded as SIGs including 36 Aesop's Fables, such as *The Fox and the Crow* in Table 3.

- The SIG framework includes an annotation tool called SCHEHERAZADE that supports representing any narrative as a SIG.

- SCHEHEREZADE comes with a realizer that regenerates stories from the SIG: this realizer provides alternative story realizations that we can compare to the EST 2.0 output.

We currently have 100 personal narratives annotated with the SIG representation on topics such as travel, storms, gardening, funerals, going to the doctor, camping, and snorkeling, selected from a corpus of a million stories (Gordon and Swanson, 2009). We use the stories in Tables 1 and 3 in this paper to explain our framework.

Fig. 2 shows the SIG for *The Startled Squirrel* story in Table 1. To create a SIG, SCHEHERAZADE annotators: (1) identify key entities; (2) model events and statives as propositions and arrange them in a timeline; and (3) model the annotator's understanding of the overarching goals, plans and beliefs of the story's agents. SCHEHERAZADE allows users to annotate a story along several dimensions, starting with the surface form of the story (first column in Table 2) and then proceeding to deeper representations. The first dimension (second column in Table 2) is called the "timeline layer", in which the story is encoded as predicate-argument structures (propositions) that are temporally ordered on a timeline. SCHEHERAZADE adapts information about predicate-argument structures from the VerbNet lexical database (Kipper et al., 2006) and uses

EST 2.0
The crow sat on the tree's branch. The cheese was in the crow's pecker. The crow thought "I will eat the cheese on the branch of the tree because the clarity of the sky is so-somewhat beautiful." The fox observed the crow. The fox thought "I will obtain the cheese from the crow's nib." The fox came. The fox stood under the tree. The fox looked toward the crow. The fox avered "I see you!" The fox alleged 'your's beauty is quite incomparable, okay?" The fox alleged 'your's feather's chromaticity is damn exquisite." The fox said "if your's voice's pleasantness is equal to your's visual aspect's loveliness you undoubtedly are every every birds's queen!" The crow thought "the fox was so-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 the wits!"

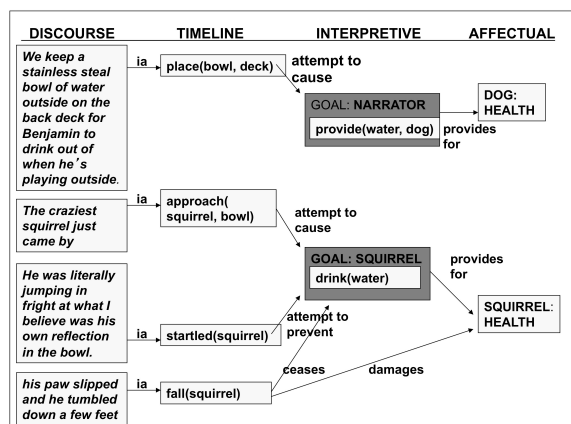Table 4: Retelling of "The Fox and the Crow"

Figure 2: Part of the STORY INTENTION GRAPH (SIG) for *The Startled Squirrel*.

WordNet (Fellbaum, 1998) as its noun and adjectives taxonomy. The arcs of the story graph are labeled with discourse relations, such as *attempts to cause*, or *temporal order* (see Chapter 4 of (Elson, 2012).)

The EST applies a model of syntax to the SIG which translates from the semantic representation of the SIG to the syntactic formalism of Deep Syntactic Structures (DSYNTS) required by the PERSONAGE generator (Lavoie and Rambow, 1997; Melčuk, 1988; Mairesse and Walker, 2011). Fig. 1 provides a high level view of the architecture of EST. The full translation methodology is described in (Rishes et al., 2013).

DSYNTS are a flexible dependency tree representation of an utterance that gives us access to the underlying linguistic structure of a sentence that goes beyond surface string manipulation. The nodes of the DSYNTS syntactic trees are labeled with lexemes and the arcs of the tree are labeled with syntactic relations. The DSYNTS formalism distinguishes between arguments and modifiers and between different types of arguments

Variation	Blog Output	Fable Output
Original	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 Crow was hugely flattered by this, and just to show the Fox that she could sing she gave a loud caw.
Sch	A narrator placed a steely and large bowl on a back deck in order for a dog to drink the water of the bowl.	The crow cawed loudly in order for she to show him that she was able to sing.
EST 1.0	I placed the bowl on the deck in order for Benjamin to drink the bowl's water.	The crow cawed loudly in order to show the fox the crow was able to sing.
becauseNS	I placed the bowl on the deck because Benjamin wanted to drink the bowl's water.	The crow cawed loudly because she wanted to show the fox the crow was able to sing.
becauseSN	Because Benjamin wanted to drink the bowl's water, I placed the bowl on the deck.	Because the crow wanted to show the fox the crow was able to sing, she cawed loudly.
NS	I placed the bowl on the deck. Benjamin wanted to drink the bowl's water.	The crow cawed loudly. She wanted to show the fox the crow was able to sing.
N	I placed the bowl on the deck.	The crow cawed loudly.
soSN	Benjamin wanted to drink the bowl's water, so I placed the bowl on the deck.	The crow wanted to show the fox the crow was able to sing, so she cawed loudly.

Table 5: Sentence Planning Variations added to EST 2.0 for Contingency relations, exemplified by *The Startled Squirrel* and *The Fox and the Crow*. Variation **N** is intended to test whether the content of the satellite can be recovered from context. **Sch** is the realization produced by Scheherezade.

(subject, direct and indirect object etc). Lexicalized nodes also contain a range of grammatical features used in generation. RealPro handles morphology, agreement and function words to produce an output string.

This paper utilizes the ability of the EST 2.0 and the flexibility of DSYNTS to produce direct speech that varies the character voice as illustrated in Table 4 (Lukin and Walker, 2015). By simply modifying the `person` parameter in the DSYNTS, we can change the sentence to be realized in the first person. For example, to produce the variations in Table 4, we use both first person, and direct speech, as well as linguistic styles from PERSONAGE: a neutral voice for the narrator, a shy voice for the crow, and a laid-back voice for the fox (Lukin and Walker, 2015). We fully utilize this variation when we retell personal narratives in EST 2.0.

This paper and introduces support for new discourse relations, such as aggregating clauses related by the contingency discourse relation (one of many listed in the Penn Discourse Tree Bank (PDTB) (Prasad et al., 2008)). In SIG encoding, contingency clauses are always expressed with the "in order to" relation (Table 6, 1). To support linguistic variation, we introduce "de-aggregation" onto these aggregating clauses in order to have the flexibility to rephrase, restructure, or ignore clauses as indicated by our parameterized sentence planner. We identify candidate story points in the SIG that contain a contingency relation (annotated in the Timeline layer) and deliberately break apart

this hard relationship to create nucleus and satellite DSYNTS that represents the entire sentence (Table 6, 2) (Mann and Thompson, 1988). We create a text plan (Table 6, 3) to allow the sentence planner to reconstruct this content in various ways. Table 5 shows sentence planning variations for the contingency relation for both fables and personal narratives (**soSN**, **becauseNS**, **becauseSN**, **NS**, **N**), the output of EST 1.0, the original sentence (**original**), and the SCHEHERAZADE realization (**Sch**) which provides an additional baseline. The **Sch** variant is the original "in order to" contingency relationship produced by the SIG annotation. The **becauseNS** operation presents the *nucleus* first, followed by a *because*, and then the *satellite*. We can also treat the nucleus and satellite as two different sentences (**NS**) or completely leave off the satellite (**N**). We believe the **N** variant is useful if the satellite can be easily inferred from the prior context.

The richness of the discourse information present in the SIG and our ability to de-aggregate and aggregate will enable us to implement other discourse relations in future work.

3 Personal Narrative Evaluation

After annotating our 100 stories with the SCHEHERAZADE annotation tool, we ran them through the EST, and examined the output. We discovered several bugs arising from variation in the blogs that are not present in the Fables, and fixed them. In previous work on the EST, the machine translation metrics Levenshtein's distance and BLEU score were used to compare

Table 6: 1: original unbroken DSYNTS; 2) deaggregated DSYNTS; 3) contingency text plan

```
1: ORIGINAL
<dsynts id="5_6">
 <dsyntnode class="verb" lexeme="organize"
   mode="" mood="ind" rel="II" tense="past">
  <dsyntnode article="def" class="common_noun"
    lexeme="bird" number="pl" person="" rel="I"/>
  <dsyntnode article="def" class="common_noun"
    lexeme="bird" number="pl" person="" rel="II"/>
  <dsyntnode class="preposition" lexeme="on"
        rel="ATTR">
   <dsyntnode article="def" class="common_noun"
     lexeme="railing" number="sg" person="" rel="II">
    <dsyntnode article="no-art" class="common_noun"
      lexeme="deck" number="sg" person="" rel="I"/>
   </dsyntnode>
  </dsyntnode>
  <dsyntnode class="preposition" lexeme="in_order"
      rel="ATTR">
   <dsyntnode class="verb" extrapo="+" lexeme="wait"
     mode="inf-to" mood="inf-to"
     rel="II" tense="inf-to">
    <dsyntnode article="def" class="common_noun"
      lexeme="bird" number="pl" person="" rel="I"/>
   </dsyntnode>
  </dsyntnode>
 </dsyntnode>
</dsynts>
```

```
2: DEAGGREGATION
<dsynts id="5">
 <dsyntnode class="verb" lexeme="organize"
   mood="ind" rel="II" tense="past">
  <dsyntnode article="def" class="common_noun"
    lexeme="bird" number="pl" person="" rel="I"/>
  <dsyntnode article="def" class="common_noun"
    lexeme="bird" number="pl" person="" rel="II"/>
  <dsyntnode class="preposition" lexeme="on"
        rel="ATTR">
   <dsyntnode article="def" class="common_noun" l
     lexeme="railing" number="sg"
     person="" rel="II">
    <dsyntnode article="no-art" class="common_noun"
      lexeme="deck" number="sg" person="" rel="I"/>
   </dsyntnode>
  </dsyntnode>
 </dsyntnode>
</dsynts>

<dsynts id="6">
 <dsyntnode class="verb" lexeme="want"
   mood="ind" rel="II" tense="past">
  <dsyntnode article="def" class="common_noun"
    lexeme="bird" number="pl" person="" r
  <dsyntnode class="verb" extrapo="+"
   lexeme="wait" mode="inf-to" mood="inf-to"
   rel="II" tense="inf-to"/>
 </dsyntnode>
</dsynts>
```

```
3: AGGREGATION TEXT PLAN
 <speechplan voice="Narrator">
 <rstplan>
  <relation name="contingency_cause">
   <proposition id="1" ns="nucleus"/>
   <proposition id="2" ns="satellite"/>
  </relation>
 </rstplan>
 <proposition dialogue_act="5" id="1"/>
 <proposition dialogue_act="6" id="2"/>
</speechplan>
```

the original Aesop's Fables to their generated EST and SCHEHERAZADE reproductions (denoted **EST** and **Sch**) (Rishes et al., 2013). These metrics are not ideal for evaluating story quality, especially when generating stylistic variations of the original story. However they allow us to automatically test some aspects of system coverage, so we repeat this evaluation on the blog dataset.

Table 7 presents BLEU and Levenshtein scores for the original 36 Fables and all 100 blog stories, compared to both **Sch** and EST 1.0. Levenshtein

distance computes the minimum edit distance between two strings, so we compare the entire original story to a generated version. A lower score indicates a closer comparison. BLEU score computes the overlap between two strings taking word order into consideration: a higher BLEU score indicates a closer match between candidate strings. Thus Table 7 provides quantitative evidence that the style of the original blogs is very different from Aesop's Fables. Neither the EST output nor the **Sch** output comes close to representing the original textual style (Blogs Original-Sch and Original-EST).

Table 7: Mean for Levenshtein and BLEU on the Fables development set vs. the Blogs

		Lev	**BLEU**
FABLES	Sch-EST	72	.32
	Original-Sch	116	.06
	Original-EST	108	.03
BLOGS	Sch-EST	110	.66
	Original-Sch	736	.21
	Original-EST	33	.21

However we find that **EST** compares favorably to **Sch** on the blogs with a relatively low Levenshtein score, and higher BLEU score (Blogs Sch-EST) than the original Fables evaluation (Fables Sch-EST). This indicates that even though the blogs have a diversity of language and style, our translation comes close to the **Sch** baseline.

4 Experimental Design and Results

We conduct two experiments on Mechanical Turk to test variations generated with the deaggregation and point of view parameters. We compare the variations amongst themselves and to the original sentence in a story. We are also interested in identifying differences among individual stories.

In the first experiment, we show an excerpt from the original story telling and indicate to the participants that "any of the following sentences could come next in the story". We then list all variations of the following sentence with the "in order to" contingency relationship (examples from the *Startled Squirrel* labeled EST 2.0 in Table 5).

Our aim is to elicit rating of the variations in terms of correctness and goodness of fit within the story context (1 is best, 5 is worst), and 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 also show

the original blog sentence and the EST 1.0 output before de-aggregation and sentence planning. We emphasize that the readers 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.

In the second experiment, we compare the original sentence with our best realization, and the realization produced by SCHEHEREZADE (**Sch**). We expect that SCHEHEREZADE will score more poorly in this instance because it cannot change point of view from third person to first person, even though its output is more fluent than EST 2.0 for many cases.

4.1 Results Experiment 1

We had 7 participants analyze each of the 16 story segments. All participants were native English speakers. Table 8 shows the means and standard deviations for correctness and preference rankings in the first experiment. We find that averaged across all stories, there is a clear order for correctness and preference: original, soSN, becauseNS, becauseSN, NS, EST, N.

We performed an ANOVA on preference and found that story has no significant effect on the results ($F(1, 15) = 0.18$, $p = 1.00$), indicating that all stories are well-formed and there are no outliers in the story selection. On the other hand, realization does have a significant effect on preference ($F(1, 6) = 33.74$, $p = 0.00$). This supports our hypothesis that the realizations are distinct from each other and there are preferences amongst them.

Fig. 3 shows the average correctness and preference for all stories. Paired t-tests show that there is a significant difference in reported correctness between **orig** and **soSN** ($p < 0.05$), but no difference between **soSN** and **becauseNS** ($p = 0.133$), or **becauseSN** ($p = 0.08$). There is a difference between **soSN** and **NS** ($p < 0.005$), as well as between the two different **because** operations and **NS** ($p < 0.05$). There are no other significant differences.

The are larger differences on the preference metric. Paired t-tests show that there is a significant difference between **orig** and **soSN** ($p < 0.0001$) and **soSN** and **becauseNS** ($p < 0.05$). There is no difference in preference between **becauseNS** and **becauseSN** ($p = 0.31$). However there is a significant difference between **soSN** and **becauseSN** ($p < 0.005$) and **becauseNS** and **NS** ($p < 0.0001$). Finally, there is significant difference between **becauseSN** and **NS** ($p < 0.005$) and **NS** and **EST** ($p < 0.005$). There is no difference between **EST** and **N** ($p = 0.375$), but there is a difference between **NS** and **N** ($p < 0.05$).
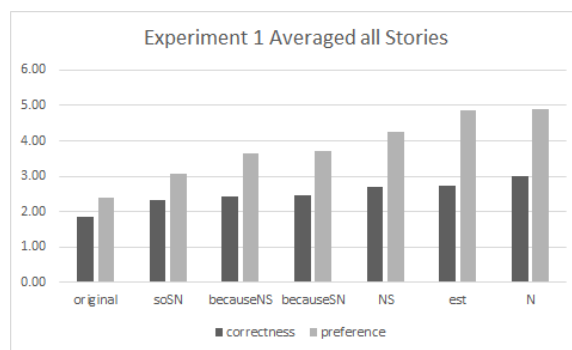

Figure 3: Histogram of Correctness and Preference for Experiment 1 averaged across story (lower is better)

These results indicate that the original sentence, as expected, is the most correct and preferred. Qualitative feedback on the original sentence included: "The one I ranked first makes a more interesting story. Most of the others would be sufficient, but boring."; "The sentence I ranked first makes more sense in the context of the story. The others tell you similar info, but do not really fit.". Some participants ranked **soSN** as their preferred variant (although the difference was never statistically significant): "The one I rated the best sounded really natural."

Although we observe an overall ranking trend, there are some differences by story for **NS** and **N**. Most of the time, these two are ranked the lowest. Some subjects observe: "#1 [**orig**] & #2 [**soSN**] had a lot of detail. #7 [**N**] did not explain what the person wanted to see" (a044 in Table 10); "The sentence I rated the worst [**N**] didn't explain why the person wanted to cook them, but it would have been an okay sentence." (a060 in Table 10); "I ranked the lower number [**N**] because they either did not contain the full thought of the subject or they added details that are to be assumed." (a044 in Table 10); "They were all fairly good sentences. The one I ranked worst [**N**] just left out why they decided to use facebook." (a042 in Table 10).

However, there is some support for **NS** and **N**. We also find that there is a significant interaction between story and realization ($F(2, 89) = 1.70$, $p = 0.00$), thus subjects' preference of the realization are based on the story they are reading. One subject commented: "#1 [**orig**] was the most descriptive about what family the person is looking for. I did like the way #3 [**NS**] was two sentences. It seemed to put a different emphasis on finding family" (a042 in Table 10). Another thought that the explanatory utterance altered the tone of the story: "The parent and the children in the story

193

		Orig	soSN	becauseNS	becauseSN	NS	EST	N
ALL	**C**	1.8	2.3	2.4	2.5	2.7	2.7	3.0
	P	2.4	3.1	3.7	3.8	4.2	4.9	4.9
Protest	**C**	4.9	2.7	2.4	3.9	2.1	2.7	2.7
	P	1.0	4.1	4.3	4.4	4.4	4.4	**2.8**
Story 042	**C**	4.2	4.2	4.3	3.8	3.7	4.2	2.7
	P	3.3	3.7	3.6	4.6	**3.1**	5	4

Table 8: Exp 1: Means for correctness **C** and preference **P** for original sentences and generated variations for **ALL** stories vs. the **Protest Story** and **a042** (stimuli in Table 10). Lower is better.

were having a good time. It doesn't make sense that parent would want to do something to annoy them [the satellite utterance]" (a060 in Table 10). This person preferred leaving off the satellite and ranked **N** as the highest preference.

We examined these interactions between story and preference ranking for **NS** and **N**. This may be depend on either context or on the SIG annotations. For example, in one story (protest in Table 10) our best realization **soSN**, produces: "The protesters wanted to block the street, so the person said for the protesters to protest in the street in order to block it." and **N** produces "The person said for the protesters to protest in the street in order to block it.". One subject, who ranked **N** second only to **original**, observed: "Since the police were coming there with tear gas, it appears the protesters had already shut things down. There is no need to tell them to block the street." Another subject who ranked **N** as second preference similarly observed "Frankly using the word protesters and protest too many times made it seem like a word puzzle or riddle. The meaning was lost in too many variations of the word 'protest.' If the wording was awkward, I tried to assign it toward the 'worst' end of the scale. If it seemed to flow more naturally, as a story would, I tried to assign it toward the 'best' end."
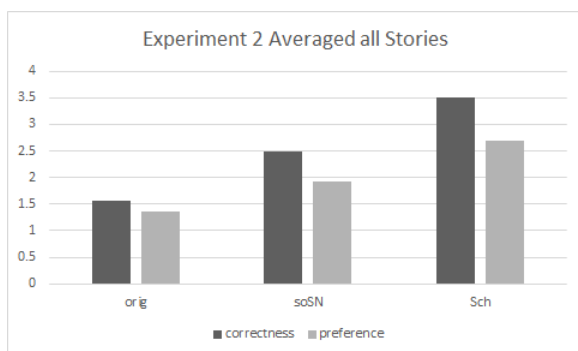

Figure 4: Histogram of Correctness and Preference for Experiment 1 averaged across story (lower is better)

Although the means in this story seem very distinct (Table 8), there is only a significant difference between **orig** and **N** ($p < 0.005$) and **N** and **EST** ($p < 0.05$). Table 8 also includes the means for story a042 (Table 10) where **NS** is ranked highest for preference. Despite this, the only significant difference between **NS** is with **EST 1.0** ($p < 0.05$).

4.2 Results Experiment 2

Experiment 2 compares our best realization to the SCHEHERAZADE realizer, exploiting the ability of EST 2.0 to change the point of view. Seven participants analyzed each of the 16 story segments. All participants were native English speakers.

	Original	soSN	Sch
Correctness	1.6	2.5	3.5
Preference	1.4	1.9	2.7

Table 9: Exp 2: Means for correctness and preference for original sentence, our best realization **soSN**, and Sch. Lower is better.

Table 9 shows the means for correctness and preference rankings. Figure 4 shows a histogram of average correctness and preference by realization for all stories. There is a clear order for correctness and preference: original, soSN, Sch, with significant differences between all pairs of realizations ($p < 0.0001$).

However, in six of the 19 stories, there is no significant difference between Sch and soSN. Three of them do not contain "I" or "the narrator" in the realization sentence. Many of the subjects comment that the realization with "the narrator" does not follow the style of the story: "The second [**Sch**] uses that awful 'narrator.'" (a001 in Table 10); "Forget the narrator sentence. From here on out it's always the worst!" (a001 in Table 10). We hypothesize that in the three sentences without "the narrator", **Sch** can be properly evaluated without the "narrator" bias. In fact, in these situations, **Sch** was rated higher than **soSN**: "I chose

the sentences in order of best explanatory detail" (*Startled Squirrel* in Table 5).

Compare the **soSN** realization in the protest story in Table 10 "The leaders wanted to talk, so they met near the workplace." with **Sch** "The group of leaders was meeting in order to talk about running a group of countries and near a workplace." **Sch** has so much more detail than **soSN**. While the EST has massively improved and overall is preferred to **Sch**, some semantic components are lost in the translation process.

5 Discussion and Conclusions

To our knowledge, this is the first time that sentence planning variations for story telling have been implemented in a framework where the discourse (telling) is completely independent of the fabula (content) of the story (Lonneker, 2005). We also show for the first time that the SCHEHEREZADE annotation tool can be applied to informal narratives such as personal narratives from weblogs, and the resulting SIG representations work with existing tools for translating from the SIG to a retelling of a story.

We present a parameterized sentence planner for story generation, that provides aggregation operations and variations in point of view. The technical aspects of de-aggregation and aggregation builds on previous work in NLG and our earlier work on SPaRKy (Cahill et al., 2001; Scott and de Souza, 1990; Paris and Scott, 1994; Nakatsu and White, 2010; Howcroft et al., 2013; Walker et al., 2007; Stent and Molina, 2009). However we are not aware of previous NLG applications needing to first de-aggregate the content, before applying aggregation operations.

Our experiments show that, as expected, readers almost always prefer the original sentence over automatically produced variations, but that the **soSN** variant is preferred. We examine two specific stories where preferences vary from the overall trend: these stories suggest future possible experiments where we might vary more aspects of the story context and audience. We also compare our best variation to what SCHEHERAZADE produces. Despite the fact that the SCHEHERAZADE realizer was targeted at the SIG, our best variant is most often ranked as a preferred choice.

In future work, we aim to explore interactions between a number of our novel narratological parameters. We expect to do this both with a rule-based approach, as well as by building on recent work on statistical models for expressive generation (Rieser and Lemon, 2011; Paiva and Evans, 2004; Langkilde, 1998; Rowe et al., 2008; Mairesse and Walker, 2011). This should allow us to train a narrative generator to achieve particular narrative effects, such as engagement or empathy with particular characters. We will also expand the discourse relations that EST 2.0 can handle.

Appendix. Table 10 provides additional examples of the output of the EST 2.0 system, illustrating particular user preferences and system strengths and weaknesses.

References

E. André, T. Rist, S. van Mulken, M. Klesen, and S. Baldes. 2000. The automated design of believable dialogues for animated presentation teams. *Embodied conversational agents*, pp. 220–255.

T.W. Bickmore. 2003. *Relational agents: Effecting change through human-computer relationships*. Ph.D. thesis, MIT Media Lab.

L. Cahill, J. Carroll, R. Evans, D. Paiva, R. Power, D. Scott, and K. van Deemter. 2001. From rags to riches: exploiting the potential of a flexible generation architecture. In *ACL-01*

C.B. Callaway and J.C. Lester. 2002. Narrative prose generation* 1. *Artificial Intelligence*, 139(2):213–252.

N. Dethlefs, H. Cuayáhuitl, H. Hastie, V. Rieser, and O. Lemon. 2014. Cluster-based prediction of user ratings for stylistic surface realisation. *EACL 2014*, page 702.

D.K. Elson and K.R. McKeown. 2009. A tool for deep semantic encoding of narrative texts. In *Proc. of the ACL-IJCNLP 2009 Software Demonstrations*, pp. 9–12.

D.K. Elson. 2012. *Modeling Narrative Discourse*. Ph.D. thesis, Columbia University, New York City.

C. Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

K. Forbes-Riley and D. Litman. 2011. Designing and evaluating a wizarded uncertainty-adaptive spoken dialogue tutoring system. *Computer Speech & Language*, 25(1):105–126.

A. Gordon and R. Swanson. 2009. Identifying personal stories in millions of weblog entries. In *Third Int. Conf. on Weblogs and Social Media, Data Challenge Workshop, San Jose, CA*.

J. Gratch, L.P. Morency, S. Scherer, G. Stratou, J. Boberg, S. Koenig, T. Adamson, A. Rizzo, et al. 2012. User-state sensing for virtual health agents and telehealth applications. *Studies in health technology and informatics*, 184:151–157.

a001	Bug out for blood the other night, I left the patio door open just long enough to let in a dozen bugs of various size. I didn't notice them until the middle of the night, when I saw them clinging to the ceiling. I grabbed the closest object within reach, and with a rolled-up comic book I smote mine enemies and smeared their greasy bug guts. All except for the biggest one. I only clipped that one, taking off one of its limbs. But it got away before I could finish the job. So now there's a five-limbed insect lurking in the apartment, no doubt looking for some vengeance against me.
orig	I'm looking around corners, checking the toilet before sitting down
Sch	The narrator began to check the toilet seat of the narrator for the leader of the group of bugs in order for she to sit down on the toilet seat of the narrator
EST	I wanted to sit down on my toilet seat, so I checked my toilet seat for the bugs's leader.
a042-1	This last week I was exploring the FamilySearch Learning Center area to see what was new. I randomly choose a video to check it out. In the middle of the video he was talking about using Facebook as a research tool. I listened as he talked about finding a family group on Facebook, on which he found pictures and information.
orig	I considered this and decided it was time to set up our Family Facebook to find those branches that have not yet been found.
Sch	The narrator decided to use the utility in order to find the family of the narrator.
soSN	I wanted to find my family, so I decided to use Facebook.
becauseNS	I decided to use Facebook because I wanted to find my family.
becauseSN	Because I wanted to find my family, I decided to use Facebook.
NS	I decided to use Facebook. I wanted to find my family.
EST	I decided to use Facebook in order for me to find my family.
N	I decided to use Facebook.
a044	I'm writing this from the Bogota airport, waiting for my flight back to Santiago. When I last posted, I was about to head off to northern Colombia to Cartagena. So, how was it?
orig	I split the 4 days between Cartagena and Santa Marta
Sch	
soSN	I wanted to see Cartagena, so I traveled to Colombia.
becauseNS	I traveled to Colombia because I wanted to see Cartagena.
becauseSN	Because I wanted to see Cartagena, I traveled to Colombia.
NS	I traveled to Colombia. I wanted to see Cartagena.
EST	I traveled to Colombia in order for me to see Cartagena and for me to see Santa Marta.
N	I traveled to Colombia.
a060-1	I hope everyone survived the snow! With the early school dismissal on Friday, it felt like a 3 day weekend. My kids are just not creatures of Winter. I did manage to take them and some of the neighborhood kids out sledding on Friday and Saturday. That was a blast. The kids had more fun, and I had a fire in the shelter with a bag of marshmallows and just enjoying myself. Followed up, of course, with hot chocolate at home. I even managed to cook cornbread from scratch, in an old (my grandmothers) cast-iron skillet, with chicken and gravy for dinner.
orig	If I had any collard greens, I think I would have cooked them too (just to annoy the kids).
Sch	The narrator wanted to cook a group of collards in order to annoy the group of children of the narrator.
soSN	I wanted to annoy my children, so I wanted to cook the collards.
becauseNS	I wanted to cook the collards because I wanted to annoy my children.
becauseSN	Because I wanted to annoy my children, I wanted to cook the collards.
NS	I wanted to cook the collards. I wanted to annoy my children.
EST	I wanted to cook the collards in order for me to annoy my child.
N	I wanted to cook the collards.
protest	The protesters apparently started their protest at the Capitol Building then moved to downtown. We happened to be standing at the corner of 16th and Stout when somebody said that the Police were getting ready to tear-gas a group of demonstrators. We looked around the corner and there were Police everywhere.
orig	They had blockaded the whole street, and shut down the light rail.
Sch	A person said that the group of protesters had protested in a street and in order to block the street.
soSN	The protesters wanted to block the street, so the person said for the protesters to protest in the street in order to block it.
becauseNS	The person said for the protesters to protest in the street in order to block it because the protesters wanted to block the street.
becauseSN	Because the protesters wanted to block the street, the person said for the protesters to protest in the street in order to block it.
NS	The person said for the protesters to protest in the street in order to block it. The protesters wanted to block the street.
EST	The person said for the protesters to protest in the street in order for the protesters to block the street.
N	The person said for the protesters to protest in the street in order to block it.

Table 10: Additional Examples of EST outputs

S.J. Green and C. DiMarco. 1993. Stylistic decision-making in natural language generation. In *Proc. of the 4th European Workshop on Natural Language Generation*.

E.H. Hovy. 1988. Planning coherent multisentential text. In *Proc. 26th Annual Meeting of the Association for Computational Linguistics*, pp. 163–169.

D. M. Howcroft, C. Nakatsu, and M. White. 2013. Enhancing the expression of contrast in the SPARKY restaurant corpus. *ENLG 2013*, page 30.

Z. Hu, M. Walker, M. Neff, and J.E. Fox Tree. 2015. Storytelling agents with personality and adaptivity. *Intelligent Virtual Agents*.

D.Z. Inkpen and G. Hirst. 2004. Near-synonym choice in natural language generation. In *Recent Advances in Natural Language Processing III*.

K. Kipper, A. Korhonen, N. Ryant, and M. Palmer. 2006. Extending verbnet with novel verb classes. In *Proc. of the 6th Int. Conf. on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy.

I. Langkilde. 1998. Forest-based statistical sentence generation. In *In Proc. of the 1st Meeting of the North American Chapter of the ACL (ANLP-NAACL 2000)*, pp. 170–177.

B. Lavoie and O. Rambow. 1997. A fast and portable realizer for text generation systems. In *Proc. of the Third Conf. on Applied Natural Language Processing, ANLP97*, pp. 265–268.

B. Lonneker. 2005. Narratological knowledge for natural language generation. In *Proc. of the 10th European Workshop on Natural Language Generation (ENLG 2005)*, pp. 91–100, Aberdeen, Scotland.

S. Lukin and M. Walker. 2015. Narrative variations in a virtual storyteller. *Intelligent Virtual Agents*.

F. Mairesse and M.A. Walker. 2011. Controlling user perceptions of linguistic style: Trainable generation of personality traits. *Computational Linguistics*.

W.C. Mann and S.A. Thompson 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*.

S.W. McQuiggan, B.W. Mott, and J.C. Lester. 2008. Modeling self-efficacy in intelligent tutoring systems: An inductive approach. *User Modeling and User-Adapted Interaction*, 18-1:81123.

I.A. Melčuk. 1988. *Dependency Syntax: Theory and Practice*. SUNY, Albany, New York.

N. Montfort. 2007. *Generating narrative variation in interactive fiction*. Ph.D. thesis, University of Pennsylvania.

C. Nakatsu and M. White. 2010. Generating with Discourse Combinatory Categorial Grammar. In *Linguistic Issues in Language Technology*, 4(1). pp. 162.

D.S. Paiva and R. Evans. 2004. A framework for stylistically controlled generation. In *Natural Language Generation, Third Int. Conf., INLG 2004*, number 3123 in LNAI, pp 120–129.

C. Paris and D. Scott. 1994. Stylistic variation in multilingual instructions. In *The 7th Int. Conf. on Natural Language Generation*.

J.W. Pennebaker and J.D. Seagal. 1999. Forming a story: The health benefits of narrative. *Journal of clinical psychology*, 55(10):1243–1254.

P. Piwek. 2003. A flexible pragmatics-driven language generator for animated agents. In *Proc. of Annual Meeting of the European Chapter of the Association for Computational Linguistics (EACL)*.

K. Porayska-Pomsta and C. Mellish. 2004. Modelling politeness in natural language generation. In *Proc. of the 3rd Conf. on INLG*, pp. 141–150.

R. Power, D. Scott, and N. Bouayad-Agha. 2003. Generating texts with style. In *Proc. of the 4th Int. Conf. on Intelligent Text Processing and Computational Linguistics*.

R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, K. Aravind B.L. Webber 2008. The Penn Discourse TreeBank 2.0. In *Language Resources and Evaluation Conference*.

V. Rieser and O. Lemon. 2011. *Reinforcement learning for adaptive dialogue systems: a data-driven methodology for dialogue management and natural language generation*. Springer.

E. Rishes, S.M. Lukin, D.K. Elson, and M.A. Walker. 2013. Generating different story tellings from semantic representations of narrative. In *Interactive Storytelling*, pp. 192–204. Springer.

J. Rowe, E. Ha, and J. Lester. 2008. Archetype-Driven Character Dialogue Generation for Interactive Narrative. In *Intelligent Virtual Agents*, pp. 45–58. Springer.

D. R. Scott and C. S. de Souza. 1990. Getting the message across in RST-based text generation. In Dale, Mellish, and Zock, ed, *Current Research in Natural Language Generation*.

A. Stent and M. Molina 2009 Evaluating automatic extraction of rules for sentence plan construction In *The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*

M.A. Walker, A. Stent, F. Mairesse and R. Prasad 2007 Individual and Domain Adaptation in Sentence Planning for Dialogue Journal of Artificial Intelligence Research (JAIR). 30:413-456.

M.A. Walker, R. Grant, J. Sawyer, G.I. Lin, N. Wardrip-Fruin, and M. Buell. 2011. Perceived or not perceived: Film character models for expressive nlg. In *Int. Conf. on Interactive Digital Storytelling, ICIDS'11*.

N. Wang, W. Lewis Johnson, R.E. Mayer, P. Rizzo, E. Shaw, and H. Collins. 2005. The politeness effect: Pedagogical agents and learning gains. *Frontiers in Artificial Intelligence and Applications*, 125:686–693.

I. Zukerman and D. Litman. 2001. Natural language processing and user modeling: Synergies and limitations. *User Modeling and User-Adapted Interaction*, 11(1-2):129–158.