Perceived or Not Perceived: Film Character Models for Expressive NLG

Marilyn A. Walker, Ricky Grant, Jennifer Sawyer,
Grace I. Lin, Noah Wardrip-Fruin, and Michael Buell

Natural Language and Dialogue Systems Lab,
University of California Santa Cruz, Santa Cruz, Ca. 95064, USA

Abstract. This paper presents a method for learning models of character linguistic style from a corpus of film dialogues and tests the method in a perceptual experiment. We apply our method in the context of SpyFeet, a prototype role playing game. In previous work, we used the PERSONAGE engine to produce restaurant recommendations that varied according to the speaker's personality. Here we show for the first time that: (1) our expressive generation engine can operate on content from the story structures of an RPG; (2) PERSONAGE parameter models can be learned from film dialogue; (3) PERSONAGE rule-based models for extraversion and neuroticism are be perceived as intended in a new domain (SpyFeet character utterances); and (4) that the parameter models learned from film dialogue are generally perceived as being similar to the character that the model is based on. This is the first step of our long term goal to create off-the-shelf tools to support authors in the creation of interesting dramatic characters and dialogue partners, for a broad range of types of interactive stories and role playing games.

Keywords: Story dialog, statistical natural language generation, character design.

1 Introduction

Stories are told through plot structure and narrative, but also critically through dialogue — what a character says, how he says it, and how he reacts to what other characters say. It is widely agreed that progress in interactive story and narrative systems is being hampered by the current approach to dialogue creation, which relies on an individual practitioner's expertise in the creative writing of dialog, often written and rewritten many times [8,16,15]. This places a hard limit on the underlying system states that can be expressed [25]. Moreover, the problem is exacerbated when authoring stories that the user is intended to experience many times, with different story trajectories depending on the user's choices and history. It has been suggested that natural language generation techniques promise to overcome the dialogue authoring bottleneck for interactive stories and games [25], but surprisingly little work has been done on language generation for story dialogue, as opposed to generating narrative descriptions [2,20,3,4,1].

Writers commonly identify two primary challenges in dialogue writing [22]. One is the challenge of revealing subtext. Good dialogue does not explicitly state character

M. Si et al. (Eds.): ICIDS 2011, LNCS 7069, pp. 109–121, 2011.

personality (e.g., *I am a friendly person*), character emotional state (e.g., *I'm feeling hesitant*), or character motivation (e.g., *I intend to flatter you*). Rather, the most important message in most good dialogue appears as subtext, either dramatized or established indirectly by what characters actually say. The second challenge is determining how characters say what they actually say, often referred to as "finding the voice" of each character. Professional writers have developed a number of practices — such as eavesdropping in public, tape recording themselves acting the part, or creating meticulously researched character backgrounds — to help them find character voices.

This paper tests an approach to automatically creating "character voices" based on a corpus-based statistical expressive language generation engine that is trained on the IMSDb corpus of film screen plays [11]. These automatically created character voices are also intended to reveal subtext about character personality and emotion. Our method consists of three components: (1) learning models of character linguistic style from film dialogue screen plays, e.g. the dialogue in Figure 1 from Quentin Tarantino's *Pulp Fiction*; (2) using the learned models to control the parameters of PERSONAGE, an expressive language generation engine [12]; and (3) experiments on human perceptions of the character utterances created using these models. We test our approach in the context of our prototype role playing game SpyFeet [18,19], a game intended to support dynamic quest selection and dialogue generation, determined by user choices and user relationships with game characters [21].

We believe this sort of corpus-based approach is a much stronger first step than, for example, asking authors to directly tune the parameters of a natural language generation engine. The expertise required to understand the parameters involved, and their interactions, is far removed from the expertise of creative writing — while authors are quite accustomed to presenting character voices through examples, or describing a character's voice as similar to another's (or a blend of familiar voices). Further, being able to explore a landscape of utterances produced through examples could also prove a powerful tool for novice (or even expert) authors who are considering possibilities for character voices. Our initial results, described here, demonstrate that an approach of this sort can produce significant and recognizable variations in linguistic style, even using corpora as small as the utterances of a single character in a screenplay.

Section 2 explains how we use a corpus of film screen plays to learn models of the linguistic style of film characters. Section 3 presents our experimental design, where we first establish human perceptions of the personality of film characters, and then test perceptions of the personality of utterances generated using both learned character models and Big Five personality models. Section 4 presents our experimental results. In previous work, we used the PERSONAGE engine to produce restaurant recommendations that varied according to the speaker's personality, where personality was defined using the Big Five theory of personality [14,12]. Here we show for the first time that: (1) our expressive generation engine can operate on content from the story structures of an RPG; (2) PERSONAGE parameter models can be learned from film dialogue; (3) PERSONAGE rule-based models for extraversion and neuroticism are perceived as intended in a new domain (SpyFeet character utterances); and (4) the parameter models learned from film dialogue are generally perceived as being similar to the modelled character.

SCENE: JACKRABBIT SLIM'S, AFTER FOOD HAS ARRIVED

VINCENT: What do you think about what happened to Antwan?
MIA: Who's Antwan?
VINCENT: Tony Rocky Horror.
MIA: He fell out of a window.
VINCENT: That's one way to say it. Another way is, he was thrown out. Another was is, he was thrown out by Marsellus. And even another way is, he was thrown out of a window by Marsellus because of you.
MIA: Is that a fact?
VINCENT: No it's not, it's just what I heard.
MIA: Who told you this?
VINCENT: They.

Mia and Vincent smile.
MIA: They talk a lot, don't they?
VINCENT: They certainly do.
MIA: Well don't be shy Vincent, what exactly did they say?

Vincent is slow to answer.
MIA: Let me help you Bashful, did it involve the F-word?
VINCENT: No. They just said Rocky Horror gave you a foot massage.
MIA: And...?
VINCENT: No and, that's it.
MIA: You heard Marsellus threw Rocky Horror out of a four-story window because he massaged my feet?
VINCENT: Yeah.
MIA: And you believed that?
VINCENT: At the time I was told, it seemed reasonable.
MIA: Marsellus throwing Tony out of a four story window for giving me a foot massage seemed reasonable?
VINCENT: No, it seemed excessive. But that doesn't mean it didn't happen. I heard Marsellus is very protective of you.
MIA: A husband being protective of his wife is one thing. A husband almost killing another man for touching his wife's feet is something else.
VINCENT: But did it happen?
MIA: The only thing Antwan ever touched of mine was my hand, when he shook it. I met Anwan once at my wedding then never again. The truth is, nobody knows why Marsellus tossed Tony Rocky Horror out of that window except Marsellus and Tony Rocky Horror. But when you scamps get together, you're worse than a sewing circle.

Fig. 1. Scene from *Pulp Fiction*

2 Learning Character Models

Procedurally generating interesting dialogue requires a large number of parameters for manipulating linguistic behavior. Our general idea is to develop corpus-based statistical models of character linguistic style by counting linguistic reflexes (features) in film dialogue, and then use these models to control the parameters of the PERSONAGE generator [14,12]. For concreteness, the PERSONAGE parameters that we wish to control with our character models are shown in Table 1. More detail is available about how PERSONAGE works elsewhere [14,12].

Corpus and Features. Our corpus consists of 862 film scripts from the IMSDb website, representing 7400 characters, with a total of 664000 lines of dialogue and 9599000 word tokens. Our snapshot of IMSDb is from May 19, 2010. We use the IMDB ontology to define groupings of character types according to the following attributes: GENRE, DIRECTOR, and YEAR. Note that most films belong to multiple genres. For example, *Pulp Fiction* belongs to crime, drama, and thriller. This allows for characters to be grouped in multiple categories. We hand-annotated CHARACTER GENDER because we thought that gender might affect linguistic style [9].

The linguistic reflexes (features) that we count in the screenplays are based on previous studies of features useful as indicators of a person's personality, gender or social

Table 1. PERSONAGE's generation parameters

Parameter	Description
Content Planning	
VERBOSITY	Control the number of propositions in the utterance
REPETITIONS	Repeat an existing proposition
CONTENT POLARITY	Control the polarity of the propositions expressed, i.e., referring to negative or positive facts
REPETITIONS POLARITY	Control the polarity of the restated propositions
CONCESSIONS	Emphasize one attribute over another
CONCESSIONS POLARITY	Determine whether positive or negative attributes are emphasized
POLARIZATION	Control whether the expressed polarity is neutral or extreme
POSITIVE CONTENT FIRST	Determine whether positive propositions — including the claim — are uttered first
INITIAL REJECTION	Begin the utterance with a mild rejection
Syntactic Template Selection	
SELF-REFERENCES	Control the number of first person pronouns
SYNTACTIC COMPLEXITY	Control the syntactic complexity (syntactic embedding)
TEMPLATE POLARITY	Control the connotation of the claim, i.e., whether positive or negative affect is expressed
Aggregation Operations	
PERIOD	Leave two propositions in their own sentences
RELATIVE CLAUSE	Aggregate propositions with a relative clause
WITH CUE WORD	Aggregate propositions using "with"
CONJUNCTION	Join two propositions using a conjunction, or a comma if more than two propositions
MERGE	Merge the subject and verb of two propositions
ALSO CUE WORD	Join two propositions using "also"
CONTRAST-CUE WORD	Contrast two propositions using "while", "but", "however", "on the other hand"
JUSTIFY-CUE WORD	Justify a proposition using "because", "since", "so"
CONCEDE-CUE WORD	Concede a proposition using "although", "even if", "but/though"
MERGE WITH COMMA	Restate a proposition by repeating only the object
Pragmatic Markers	
STUTTERING	Duplicate the first letters of a restaurant's name
PRONOMINALIZATION	Replace occurrences of the restaurant's name by pronouns
NEGATION	Negate a verb by replacing its modifier by its antonym
SOFTENER HEDGES	Insert syntactic elements to mitigate the strength of a proposition
EMPHASIZER HEDGES	Insert syntactic elements to strengthen a proposition
ACKNOWLEDGEMENTS	Insert an initial back-channel
FILLED PAUSES	Insert syntactic elements expressing hesitancy
EXCLAMATION	Insert an exclamation mark
EXPLETIVES	Insert a swear word
NEAR EXPLETIVES	Insert a near-swear word
TAG QUESTION	Insert a tag question
IN-GROUP MARKER	Refer to the hearer as a member of the same social group
Lexical Choice	
LEXICON FREQUENCY	Control the average frequency of use of each content word, according to BNC frequency counts
LEXICON WORD LENGTH	Control the average number of letters of each content word
VERB STRENGTH	Control the strength of the verbs

class [13,6,17,9]. Table 2 enumerates our feature sets. For most features, there is a particular parameter in PERSONAGE (in Table 1) whose parameter value should be affected by that feature's presence or absence in a character's dialogic utterances. The **Basic** features capture aspects of style such as how much a character talks and how many words they use (the VERBOSITY parameter). The **Dialogue Act** features are based on a dialogue act tagger trained on the NPS Chat Corpus 1.0 [5]. The **First Dialogue Act** is the Dialogue Act of the first sentence of each turn. Several dialogue act features indicate the use of the parameters INITIAL REJECTION or ACKNOWLEDGMENT. Others we do not currently utilize. **Pragmatic Markers** include both categories of pragmatic markers and individual word count/ratio. Pragmatic marker features indicate which aggregation

Table 2. Feature Sets ordered by PERSONAGE modules

Feature	Description (Label)
Basic	number of sentences (NumSents), sentences per turn (NumSentsPerTurn), number of verbs (NumVB), number of verbs per sentence (VBPerSents)
Polarity	overall polarity (polarity-overall), polarity of sentences (polarity-sents)
Dialogue Act (DA)	Accept, Bye, Clarify, Continuer, Emotion, Emphasis, Greet, No-Answer, Reject, Statement, Wh-Question, Yes-Answer, Yes-No-Question, Other
First DA	Same as DA but only look at first sentence of each turn
Merge Ratio	merging of subject and verb of two propositions (merge-ratio)
Passive Sentence Ratio	passive sentence count (passive-ratio)
Concession polarity	polarity for concessions (concess-polarity)
LIWC Word Categories	Each prefixed as LIWC-
Pragmatic Markers	wc-taboo, wc-seq, wc-opinion, wc-aggregation, wc-softeners, wc-emphatics, wc-ack, wc-pauses, wc-concession, wc-concede, wc-justify, wc-contrast, wc-conjunction, wc-ingroup, wc-near_swear, wc-relative
Tag Question Ratio	tag question ratio (tag-ratio)
Word Length	average content word length (avg-content-wlen)
Verb Strength	average sentiment values of verbs (verb-strength)

operations to use such as JUSTIFY-CUE WORD (See Table 1) or which pragmatic markers to insert, such as EMPHASIZERS or SOFTENER HEDGES. The **Merge Ratio** uses a grammar operating on part of speech labels that looks for verb+noun+conjunction+noun. The **Passive Sentence Ratio** uses scripts from http://code.google.com/p/narorumo, under source/browse/trunk/passive to detect passive sentences. These scripts implement the rule that if a to-be verb is followed by a non-gerund, the sentence is probably in passive voice. The **Concession Polarity** feature is based on finding the polarity for the concession in a sentence if it exists, using the Polarity feature set. The **LIWC** tool provides a lexical hierarchy that counts the use of different types of words, including cue-words, emotion words, and pronouns *inter alia*. These map to both aggregation operations and pragmatic markers. The **Tag Question Ratio** is also based on a set of regular expressions. The features **Word Length** and **Verb Strength** control the lexical choice parameters. **Word Length** first uses WordNet tags to find content words (noun, adjective, adverb, and verb), and then takes the mean of their length in characters. **Verb Strength** is the mean sentiment scores of all verbs. Lexical frequency is approximated from combining the features LIWC-6LTR and word length.

Method. Fig. 2 shows the flow of our experiment. In sum, our method is:

1. Collect movie scripts from The Internet Movie Script Database (IMSDb).
2. Parse each movie script to extract dialogic utterances, producing an output file containing utterances of exactly one character of each movie (e.g., *pulp-fiction-vincent.txt* has all of the lines of the character Vincent).
3. Select characters from those with more than 60 turns of dialogue.
4. Extract features representing the linguistic behaviors of each character.
5. Learn models of character linguistic styles based on the features.
6. Use character models to control parameters of the PERSONAGE generator.
7. Evaluate human perceptions of dialogic utterances generated using the character models.

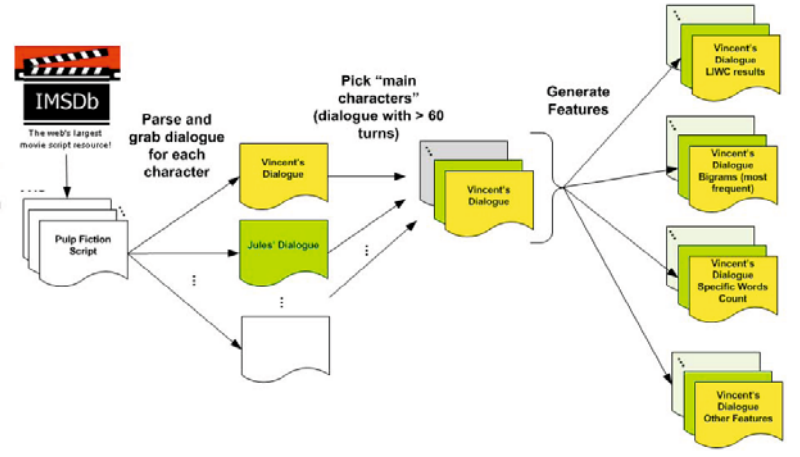

Fig. 2. Method

Character Models. Sample character models derived from the procedure above are provided in Table 3. Each model parameter in the left-hand side of Table 3 was described in Table 1. Table 4 illustrates the result of applying these models of character to SpyFeet utterances, and shows some of the variation that we are currently able to produce. For example, the Annie Hall characters, Alvy and Anny, have significant Z-scores(2.12 and 3.28 respectively) for the tag question ratio feature. The tag question ratio represents the placement of phrases like *you know?* and *would you be?* at the end of sentences. The feature value maps to a value of 1.0 for the PERSONAGE tag question insertion parameter, causing utterances generated using the Annie or Alvy character models to include the use of tag questions. The Annie and Alvy models also lead to significant z-scores for the LIWC-WC feature. LIWC-WC is is the word count for a character and maps to the verbosity parameter in PERSONAGE. The significant z-score value for LIWC-WC causes an increase in the verbosity parameter for the Alvy and Annie models, and as a result, utterances generated using these models have more words than those from models with lower verbosity values such as Vincent or Indy.

There are many different ways we could learn such models [10,23,24]. Here, we estimate models using vectors of features representing individual characters, and then derive distinctive features for that character by normalizing the feature counts against a representative population. For each feature x_i, the normalized value z_i is calculated as:

$$\frac{x_i - \overline{x_i}}{\sigma_{x_i}} \tag{1}$$

There is a choice about the population of characters used for the normalization, i.e. which set of characters are used to calculate the mean $\overline{x_i}$ and the standard deviation σ_{x_i}. For example, for a female character, obvious choices include all the characters, all the female characters, or all the female action characters. Here we normalize individual characters against all of the characters of the same gender. Any z-score greater than 1 or less than -1 is more than one standard deviation away from the mean. Z-scores greater

Table 3. Sample Learned Character Models

Parameter	Alvy	Annie	Indy	Marion	Mia	Vincent
Content Planning						
Verbosity	.79	.78	.36	.65	.49	.18
Repetitions	.38	0	0	0	.28	.51
Content Polarity	.09	.77	.15	.15	.15	.50
Polarization	.39	.72	.22	. 21	.22	.57
Repetitions Polarity	.54	.79	.29	.29	.29	.64
Concessions	.83	.83	.83	.89	.89	.58
Concessions Polarity	.56	.26	.56	.26	.26	.49
Positive Content First	0	1.00	0	0	0	1.00
Initial Rejection	0	0	0	0	0	0
Syntactic Template Selection						
Use of First Person in Claim	.39	.6	.39	.39	.39	.54
Claim Polarity	.57	.57	.57	.49	.56	.50
Claim Complexity	.71	.31	.47	.15	.56	.56
Aggregation Operations						
Period	.05	.04	.24	.04	.24	0
Relative Clause	0	0	.95	.97	.53	.3
With cue word	.44	.51	.05	.34	.31	.25
Conjunction	.30	.21	.22	.18	.08	0
Merge	.61	.87	.83	.65	.59	.77
Also cue Word	.12	.05	.05	.05	.07	.05
Contrast-Cue word	.76	.85	0	.84	.76	.96
Justify-Cue Word	.97	.48	0	.61	.61	.45
Concede-Cue Word	1.00	0	0	1.00	0	.25
Merge With Comma	.27	.42	.5	.5	.32	.5
Pragmatic Markers						
Stuttering	.54	.54	.04	.04	.54	.09
Pronominalization	1.00	1.00	1.00	.75	.5	1.00
Negation	0	0	0	0	0	0
Softener Hedges	1.00	1.00	0	1.00	0	0
Emphasizer hedges	0	1. 0	0	0	1.00	0
Acknowledgements	1.00	1.00	0	0	1.00	0
Filled Pauses	1.00	1.00	0	0	0	0
Exclamation	0	0	0	1.00	0	1.00
Expletives	0	0	0	0	0	1.00
Near Expletives	0	0	0	0	0	0
Tag Question	1.00	1.00	0	0	1.00	0
In-Group Marker	0	0	0	1.00	0	0
Lexical Choice						
Lexicon Frequency	.19	.19	.28	.19	.55	.18
Lexicon Word Length	.21	.13	.21	.13	.78	.28
Verb Strength	.59	.59	.5	.61	.5	.49

and less than +-1.96 indicate significant differences of the use of that linguistic feature by that character compared to other characters. However for experimental purposes we map any z-score greater than 1 or less than -1 into one or more PERSONAGE generation parameters. There are three different ways in which our mappings of feature counts to parameters could be incomplete: (1) We have a parameter in PERSONAGE but none of the features we count are good indicators that we should use it (e.g., Competence Mitigation); (2) We have a feature we count but no parameter to map it to (e.g., LIWC-Discrep); or (3) There is some aspect of linguistic style that is essential to expressing a particular character's style, but we currently do not have a feature that indicates when a character has that style, nor do we have any existing parameter that could manifest that linguistic reflex.

Table 4. Utterances for SpyFeet generated using Film Character Models

Film	Character	Generated Utterances
Annie Hall	**Alvy**	I don't know. People say Cartmill is st-strange, alright? Err... on the other hand, I don't rush to judgment. Right, I am not sure, would you be? I will tell something you because you br-brought me cabbage. Oh I am not sure. Wolf wears a hard shell. On the other hand, he is ge-ge-gentle, isn't he? I see, I don't know. I respect Wolf, wouldn't you? He, however, isn't my close friend. Yeah, I don't know. Sparrow conveys excitement to my life, so I am fr-fr-friends with her.
Indiana Jones	**Indy**	I don't rush to judgment, but people say Cartmill is strange. I will tell something you since you brought me cabbage. Wolf is gentle but he wears a hard shell. Wolf isn't my close friend. But I respect him. I am friends with Sparrow since she brings excitement to my life.
Annie Hall	**Annie**	Come on, I don't know, do you? People say Cartmill is strange while I don't rush to um.. judgment. I don't know. I think that you brought me cabbage, so I will tell something to you, alright? Yeah, I am not sure, would you be? Wolf wears a hard shell but he is really gentle. I see, I am not sure. Obviously, I respect Wolf. However, he isn't my close friend, is he? Come on, I am not sure. Because Sparrow brings excitement to my life, I am friends with her, you see?
Pulp Fiction	**Vincent**	Basically, I don't rush to judgment. On the other hand, people say Cartmill is strange, he is strange. Yeah, I can answer since you brought me cabbage that. Everybody knows that Wolf wears a hard shell. He, however, is gentle. I respect Wolf. However, he isn't my damn close friend. Oh God I am friends with Sparrow because she brings excitement to my life.

3 Experimental Setup

Our goal is to test the character models and mappings as described above. The simplest way to do this is to ask human participants to rate a set of utterances produced using different models in terms of similarity of linguistic style to the mimicked character. However our concern was that a single linguistic cue could act as a "give-away" for the intended character. For example, if the Annie character from Woody Allen's *Annie Hall* uses tag-questions (among other linguistic style differentiators), perhaps the use of tag-questions *alone* in a test utterance would cue a human participant that the test utterance was intended by the experimenters to mimic Annie, especially if the human participant was only asked to select between two different examples.

Therefore, we designed an experiment to first indirectly test the similarity in perceived personality of, e.g. the Annie character as written in the film, to the personality of utterances of SpyFeet characters produced using an Annie model of linguistic style. Our experimental method consists of three phases each intended to establish human perceptions of different aspects of utterances generated using character models. In **Phase I**, we select 3 scenes from each of the original films, illustrating the utterance styles of 6 characters (3 male and 3 female). We collect perceptions of the personality of those characters using the Ten Item Personality Inventory (TIPI) [7,12]. In **Phase II**, using the PERSONAGE generator, we generate dialogic utterances for the characters in the story of the SpyFeet RPG, using both (1) the film character model; and (2) six rule-based personality models from our previous work (high and low values for extraversion, neuroticism and agreeableness) [12]. We collect perceptions of the personality of SpyFeet characters whose linguistic style is controlled by these models (6 film character models and 6 personality models), again using the TIPI [7]. In **Phase III**, using all the utterances generated in Phase II, for each film character model, we generate a page showing the participant (1) the three scenes for each character (from Phase I); and (2) **all** of the generated utterances using all of the film character models and all of the rule-based personality models. Then we ask participants to judge on a scale of 1...7 how **similar** the generated utterance is to the style of the film character as illustrated in the three scenes. Participants are instructed to use the whole scale, and thus effectively **rank** the generated utterances for similarity to the film character. Each phase supports different analyses of the perceptions of SpyFeet characters. Using the data collected in Phase I, we establish participant perceptions of film characters on the Big Five personality traits of extraversion, neuroticism and agreeableness. Then using the data from Phases I and II, we examine the correlations between perceptions of the film character's original utterances (Annie, Alvy, Vincent, Mia, Indiana, Marion) and SpyFeet utterances that were generated using the learned models of the film character. Our **Hypotheses** are:

- H1: The rule-based models for personality expression (previously tested in the restaurant recommendation domain), will be perceived as expressing that personality in the SpyFeet story domain (Phase II).
- H2: Utterances generated using character models will be perceived as being more similar to that character than utterances generated using another randomly selected character model (Phase III).

4 Experimental Results

29 subjects (13 female and 16 male, ages ranging from 22 to 44) participated in a web-based experiment.

Phase I. We made no predictions about the results in Phase I. Our goal was to establish personality judgements for the six characters and test whether, in terms of Big Five traits, the characters are perceived as having distinctive personalities. Table 5 shows the mean values of the TIPI scale judgements for Big Five traits of Extraversion, Emotional Stability and Agreeableness for the six characters.

We combined the personality judgments for a character for all three Big Five traits into a single vector and computed paired t-tests (two-tailed) on these vectors to

Table 5. Big Five Personality Scores for Film Character Original Utterances

	Character					
Trait	Alvy	Annie	Indy	Marion	Mia	Vincent
Extraversion	2.8	4.4	4.2	5.5	4.8	4.6
Emotional Stability	2.0	2.5	5.0	3.8	4.4	4.1
Agreeableness	4.0	4.5	3.3	3.9	4.0	4.1

determine whether characters were perceived as having distinct personalities (within subjects). The results indicate that the personality of Alvy is perceived as being significantly different from all of the other characters (df = 90, $3.3 < t < 7.4$, $p < .001$). However Indy is only different from Alvy (df = 90, t = 4.8, $p < .0001$). Marion is only different from Alvy and Annie (df = 90, $3.3 < t < 7.4$, $p < .001$), Vincent is only different from Alvy and Annie (df = 90, $2.3 < t < 6.6$, $p < .02$), and Mia is only different from Alvy and Annie (df = 90, $3.1 < t < 7.4$, $p < .003$). These results suggest that the differences in perceived personality across different characters are small, with the Tarantino and Spielberg characters being perceived as having similar personalities. However as we show below in Phase III, there are distinctive differences in their linguistic styles that are perceivable. Our conclusion is that Big Five traits may be too coarse to effectively distinguish different film characters.

Phase II. Our prediction (Hypothesis H1) was that rule-based models for personality expression will be perceived as expressing that personality in the SpyFeet story domain. Our results for H1 are mixed. We tested whether utterances generated with high and low extraversion models, high and low agreeableness models and high and low emotional stability models are perceived as expressing those traits. A paired t-test comparing the extraversion ratings of high ($\overline{x}$ = 5.2) and low extraversion ($\overline{x}$ = 3.3) utterances showed significant differences (df = 28, t = 7.7, $p < .0001$), as did a paired t-test comparing the emotional stability ratings of high ($\overline{x}$ = 5.5) and low ($\overline{x}$ = 2.7) emotional stability utterances (df = 28, t = 10.8, $p < .0001$). However differences in high ($\overline{x}$ = 3.4) and low ($\overline{x}$ = 3.4) agreeableness were not perceived in the SpyFeet domain, when we used the agreeableness model that had previously been successful in the restaurant recommendation domain (df = 28, t = .72, p = .47 ns). There are several possible reasons for this: perhaps the limited set of utterances tested, as shown in Table 4, do not do a good job of showing the variability in agreeableness that the PERSONAGE generator is capable of, or perhaps manifesting agreeableness in the SpyFeet domain requires the addition of new parameters to PERSONAGE, or perhaps our mapping from features to parameters is either incomplete or faulty.

Phase III. Our prediction (Hypothesis H2) was that utterances generated using character models would be more similar to that character than utterances generated using another randomly selected character model. Table 6 shows the average similarity score judgments between utterances produced with a particular character model and the utterances of that character in the original film. For example Row 1 shows the judgments

for the similarity of utterances generated with each character model to the utterances of the Alvy character in the original *Annie Hall* screen play. Similarity scores are scalar values from 1...7. The strongest possible result would be a diagonal matrix with 7's along the diagonal and 0's in all the other cells, i.e. a only utterances generated with a particular character's model would be judged as being at all similar to that character. In general, what we are looking for is a matrix with the highest values along the diagonal.

Table 6. Mean Similarity Scores between Characters and Character Models. Significant differences between the designated character and each other character are shown in **bold**.

Character	Alvy	Annie	Indy	Marion	Mia	Vincent
Alvy	5.2	**4.2**	**2.1**	**2.6**	**2.8**	**2.3**
Annie	4.2	4.3	**2.8**	**3.4**	3.9	**2.9**
Indy	**1.4**	**2.2**	4.5	**2.8**	**3.3**	**3.8**
Marion	**1.6**	**2.8**	3.7	3.1	**4.1**	**4.2**
Mia	**1.7**	**2.4**	4.3	3.2	3.6	4.3
Vincent	**2.1**	**3.2**	4.5	**3.5**	**3.6**	4.6

We conducted paired t-tests comparing the similarity scores of each other character model to the similarity scores for the matching model (e.g. we compared similarity scores for utterances generated using Alvy's model to utterances generated using Indy's model, collected in the context of the participant looking at the screenplay for *Indiana Jones*.).

For *Annie Hall*, utterances generated using the Alvy model (first row of Table 6) are significantly more similar to Alvy than utterances generated using any other model ($df = 28$, $3.16 < t\ 8.35$, $p < .004$). The utterances generated using the Annie model (first row of Table 6) are significantly more similar to Annie than utterances generated with the Indy ($df = 28$, $t = 3.75$, $p < .0008$), Marion ($df = 28$, $t = 2.08$, $p < .05$), and Vincent ($df = 28$, $t = 2.90$, $p < .007$), but not different than utterances generated with the models for Alvy ($df = 28$, $t = .09$,ns), and Mia ($df = 28$, $t = .85$, ns).

For *Indiana Jones*, utterances generated using the Indy model (third row of Table 6) are significantly more similar to Indy than utterances generated using any other model ($df = 28$, $2.67 < t < 7.99$, $p < .01$) Utterances generated using the Marion model (fourth row of Table 6) are also significantly more similar to Marion than utterances generated using Alvy ($df = 28$, $t = 4.70$, $p < .0001$), Mia ($df = 28$, $t = 2.66$, $p < .013$), or Vincent models ($df = 28$, $t = 3.24$, $p < .003$), but not different than the Annie model ($df = 28$, $t = .52$, $p = .65$ ns) or the Indy model ($df = 28$, $t = 1.98$, $p < .057$).

For *Pulp Fiction*, utterances generated using the Mia model (fifth row of Table 6) are significantly more similar to Mia than utterances generated from the Alvy ($df = 28$, $t = 6.72$, $p < .0001$), and Annie ($df = 28$, $t = 3.24$, $p < .003$) models, but not different than those using models for Indy ($df = 28$, $t = 1.67$, $p = .11$ ns), Marion ($df = 28$, $t = 1.06$, $p = .30$ ns), and Vincent ($df = 28$, $t = 1.58$, $p = .13$ ns). The fact that the model for the Mia character was trained on the fewest number of utterances (she has only 81 lines in the film) could contribute to the lack of perceivable differences. Utterances generated using the Vincent model (sixth row of Table 6) are significantly more similar to Vincent

than utterances generated using Alvy (df = 28, t = 6.59 , $p < .0001$), Annie (df = 28, t = 3.54, $p < .0014$), Marion (df = 28, t = 2.57, $p < .02$), and Mia models (df = 28, t = 2.25, $p < .03$), but not different than the Indy model (df = 28, t = .86 , p = .18 ns).

5 Discussion

If deeply interactive stories are to feature dialog, we must move beyond a model of pure hand authoring. As stories vary in terms of the events that take place, the characters that are present, the dynamic states of relationships between characters, and so on, we must be able to dynamically generate dialogue that reflects and drives the state of the fictional world while expressing character in a manner controllable by an author. But asking authors to, for example, specify the parameter settings for a complex natural language generation engine is at odds with the skillsets and approaches of most authors, whether experts or beginners.

In this paper we have demonstrated the first step toward an alternative approach: developing models of character linguistic style from examples, specifically using character utterances in film scripts. Our results are encouraging, showing that utterances generated in a different domain (that of an outdoor role-playing game) recognizably display important subtext for character personality as well as style that is more similar to the modeled character than to others (though, perhaps unsurprisingly, characters from the same genre or film are often more similar to each other than to others).

After this initial step, much work remains to be done. For example, just as a character's plot actions in an interactive story must be related to the current state of the world and actions of other characters, so must linguistic actions take place in context. Our current model does not represent anything about the relation between dialogic utterances across speakers. The importance of such relations can be seen in Figure 1, in which paraphrastic and echoic aspects of the dialogue actually seem to be an interesting part of Mia's linguistic style — as well as an indication of her character's current stance toward Vincent. This points to another important area for future work, as we explore how character linguistic style varies across situations in order to help communicate emotional dynamics to the audience.

References

1. André, E., Rist, T., van Mulken, S., Klesen, M., Baldes, S.: The automated design of believable dialogues for animated presentation teams. In: Prevost, S., Cassell, J.S., Churchill, E. (eds.) Embodied Conversational Agents, pp. 220–255. MIT Press, Cambridge (2000)
2. Beskow, J., Cerrato, L., Granström, B., House, D., Nordenberg, M., Nordstrand, M., Svanfeldt, G.: Expressive Animated Agents for Affective Dialogue Systems. In: André, E., Dybkjær, L., Minker, W., Heisterkamp, P. (eds.) ADS 2004. LNCS (LNAI), vol. 3068, pp. 240–243. Springer, Heidelberg (2004)
3. Cavazza, M., Charles, F.: Dialogue generation in character-based interactive storytelling. In: AAAI First Annual Artificial Intelligence and Interactive Digital Entertainment Conference, Marina del Rey, California, USA (2005)
4. Elson, D., McKeown, K.: Automatic attribution of quoted speech in literary narrative. In: Proceedings of AAAI (2010)

5. Forsyth, E., Martell, C.: Lexical and discourse analysis of online chat dialog. IEEE Computer Society (2007)
6. Furnham, A.: Language and personality. In: Giles, H., Robinson, W. (eds.) Handbook of Language and Social Psychology, Winley (1990)
7. Gosling, S.D., Rentfrow, P.J., Swann, W.B.: A very brief measure of the big five personality domains. Journal of Research in Personality 37, 504–528 (2003)
8. Hayes-Roth, B., Brownston, L., Sincoff, E.: Directed improvisation by computer characters. Tech. Rep. KSL 95-04, Knowledge Systems Laboratory, Stanford University (1995)
9. Ireland, M., Pennebaker, J.: Authors' gender predicts their characters' language (in submission, 2011)
10. Isard, A., Brockmann, C., Oberlander, J.: Individuality and alignment in generated dialogues. In: Proceedings of the 4th International Natural Language Generation Conference (INLG 2006), Sydney, Australia, pp. 22–29 (2006)
11. Lin, G.I., Walker, M.A.: All the world's a stage: Learning character models from film. In: Proceedings of the Seventh AI and Interactive Digital Entertainment Conference, AIIDE 2011. AAAI (2011)
12. Mairesse, F., Walker, M.: Towards personality-based user adaptation: psychologically informed stylistic language generation. User Modeling and User-Adapted Interaction, 1–52 (2010)
13. Mairesse, F., Walker, M.A., Mehl, M.R., Moore, R.K.: Using linguistic cues for the automatic recognition of personality in conversation and text. Journal of Artificial Intelligence Research (JAIR) 30, 457–500 (2007)
14. Mairesse, F., Walker, M.A.: Controlling user perceptions of linguistic style: Trainable generation of personality traits. Computational Linguistics (2011)
15. Mateas, M.: The authoring bottleneck in creating AI-based interactive stories. In: Proceedings of the AAAI 2007 Fall Symposium on Intelligent Narrative Technologies (2007)
16. Mateas, M., Stern, A.: Façade: An experiment in building a fully-realized interactive drama. In: Proceedings of the Game Developers Conference, Game Design Track (2003)
17. Pennebaker, J.W., King, L.A.: Linguistic styles: Language use as an individual difference. Journal of Personality and Social Psychology 77, 1296–1312 (1999)
18. Reed, A., Samuel, B., Sullivan, A., Grant, R., Grow, A., Lazaro, J., Mahal, J., Kurniawan, S., Walker, M., Wardrip-Fruin, N.: Spyfeet: An exercise rpg. In: Proceedings of the Sixth International Conference on the Foundations of Digital Games, FDG 2011, pp. 310–312. ACM, New York (2011)
19. Reed, A., Samuel, B., Sullivan, A., Grant, R., Grow, A., Lazaro, J., Mahal, J., Kurniawan, S., Walker, M., Wardrip-Fruin, N.: A step towards the future of role-playing games: The spyfeet mobile rpg project. In: Proceedings of the Seventh AI and Interactive Digital Entertainment Conference, AIIDE 2011. AAAI (2011)
20. Rowe, J.P., Ha, E.Y., Lester, J.C.: Archetype-Driven Character Dialogue Generation for Interactive Narrative. In: Prendinger, H., Lester, J.C., Ishizuka, M. (eds.) IVA 2008. LNCS (LNAI), vol. 5208, pp. 45–58. Springer, Heidelberg (2008)
21. Sullivan, A., Mateas, M., Wardrip-Fruin, N.: Rules of engagement: moving beyond combat-based quests. In: Proceedings of the Intelligent Narrative Technologies III Workshop, pp. 1–8. ACM (2010)
22. Trottier, D.: The Screenwriter's Bible: A Complete Guide to Writing, Formatting, and Selling Your Script, 5th edn. Silman-James Press (2010)
23. Walker, M., Rambow, O., Rogati, M.: Training a sentence planner for spoken dialogue using boosting. Computer Speech and Language 16(3-4) (2002)
24. Walker, M.A., Stent, A., Mairesse, F., Prasad, R.: Individual and domain adaptation in sentence planning for dialogue. Journal of Artificial Intelligence Research (JAIR) 30, 413–456 (2007)
25. Wardrip-Fruin, N.: Expressive Processing: Digital fictions, computer games, and software studies. The MIT Press (2009)