Murder in the Arboretum: Comparing Character Models to Personality Models

Marilyn A. Walker, Grace I. Lin, Jennifer Sawyer,
Ricky Grant, Michael Buell, Noah Wardrip-Fruin
Natural Language and Dialogue Systems Lab, Computer Science Dept.
University of California, Santa Cruz
1156 High Street
Santa Cruz, California 95064
maw—glin—jsawyer—rgrant—mbuell—nwf@soe.ucsc.edu
http://nlds.soe.ucsc.edu

Abstract

Interactive Narrative often involves dialogue with virtual dramatic characters. In this paper we compare two kinds of models of character style: one based on models derived from the Big Five theory personality, and the other derived from a corpus-based method applied to characters and films from the IMSDb archive. We apply these models to character utterances for a pilot narrative-based outdoor augmented reality game called *Murder in the Arboretum*. We use an objective quantitative metric to estimate the quality of a character model, with the aim of predicting model quality without perceptual experiments. We show that corpus-based character models derived from individual characters are often more detailed and specific than personality based models, but that there is a strong correlation between personality judgments of original character dialogue and personality judgments of utterances generated for *Murder in the Arboretum* that use the derived character models.

Introduction

Conversation is an essential component of social behavior, one of the primary means by which humans express emotions, moods, attitudes and personality. Thus a key technical capability for interactive narrative systems (INS) is the ability to support natural conversational interaction. To do so, natural language processing can be used to process the user's input to allow users flexibility in what they say to the system (Johnson et al. 2005; Mateas and Stern 2003; Louchart et al. 2005). However, in most interactive narrative systems to date, character dialogue is highly handcrafted. Although this approach offers total authorial control and produces high quality utterances, it suffers from problems of portability and scalability (Walker and Rambow 2002), or what has been called the *authoring bottleneck* (Mateas 2007; Short 2009). Moreover, handcrafting makes it difficult, if not impossible, to personalize the dialogue interaction, but personalization leads to perceptions of greater player agency (Murray 1997; Hayes-Roth and Brownston 1994; Mott and Lester 2006; Thue, Bulitko, and Spetch 2008).

(Laszlo and Ilsa enter Rick's Cafe)
Headwaiter: Yes, M'sieur?
Laszlo: I reserved a table. Victor Laszlo.
Waiter: Yes, M'sieur Laszlo. Right this way.
(Laszlo and Ilsa follow the waiter to a table)
Laszlo: Two cointreaux, please.
Waiter: Yes, M'sieur.

Figure 1: Excerpt from the *Casablanca* script.

(1) Bring us two cointreaux, right away.
(2) You must bring us two cointreaux.
(3) We don't have two cointreaux, yet.
(4) You wouldn't want to bring us two cointreaux, would you?

Figure 2: Generated variations of Laszlo's request

Expressive Natural Language Generation (ENLG) promises a solution to these problems, but the ENLG engine must be able to produce variations in linguistic style that clearly manifest differences in dramatic character. To see how linguistic style conveys dramatic character, consider Laszlo's request for two cointreaux in the *Casablanca* screenplay in Fig. 1, and the automatically generated variants of that request in Fig. 2 (Walker, Cahn, and Whittaker 1997a). Clearly, speakers make stylistic choices when they realize their communicative intentions, and their realizations express their character and personality. Moreover, variations in linguistic style are the basis for listener inferences about the speaker (Cassell and Bickmore 2003; Wang et al. 2005; Rapp, Gerrig, and Prentice 2001; Brennan and Ohaeri 1994; Ross 1977; Isbister and Nass 2000). For example, someone listening to Utterance (4) in Fig. 2 might infer, given the situational context, that Laszlo is a rather wimpy hero.

Therefore the first requirement for building an ENLG for dialogue for dramatic characters, is a method or a theory that systematically and comprehensively quantifies the most important individual and stylistic differences in behavior, the way they affect linguistic output in dialogue, and the predicted effect on the perceptions of the listener. Previous work on ENLG has explored parameters and models based on Brown and Levinson's theory of politeness, the Big Five theory of personality, and dramatic theories of archetypes, (Piwek 2003; André et al. 2000; Mairesse and Walker 2010;

header:	REQUEST-ACT(speaker, hearer, action)
precondition:	WANT(speaker,action)
	CANDO(hearer,action)
decomposition-1:	surface-request(speaker,hearer,action)
decomposition-2:	surface-request(speaker,hearer, INFORMIF (hearer,speaker, CANDO(hearer,action)))
decomposition-3:	surface-inform(speaker,hearer, ¬(CANDO(speaker,action)))
decomposition-4:	surface-inform(speaker,hearer, WANT(speaker,action))
effects:	WANT(hearer,action)
	KNOW(hearer, WANT(speaker, action))
constraint:	AGENT (action,hearer)

Figure 3: Definition of the REQUEST-ACT plan operator from Litman and Allen, 1990

header:	SERVE(waiter, customer, two-cointreaux)
precondition:	HAS(restaurant, two-cointreaux)
decomposition:	BRING(waiter, customer, two-cointreaux)
effects:	HAS(customer, two-cointreaux)

Figure 4: A possible plan in the restaurant domain for serving two cointreaux

Gupta, Walker, and Romano 2007; Walker, Cahn, and Whittaker 1997a; Wang et al. 2005; Rowe, Ha, and Lester 2008; Cavazza and Charles 2005) *inter alia*. Here we compare and contrast two different bases for character models of linguistic style: one based on Big Five personality and the other a corpus-based approach that incorporates some concepts from the dramatic theory of archetypes.

We first lay out our assumptions about how interactive stories must be represented to support ENLG. Then we describe two theoretical and empirical bases for linguistic style and compare and contrast them. We then describe two kinds of experiments comparing corpus-based and personality-based models of linguistic style. We first objectively compare a corpus-based approach that utilizes single characters, to an approach that attempts to combine characters based on the dramatic archetype that they instantiate, in terms of how detailed the models are and how many parameters they specify. We show that, to date, models based on dramatic archetypes appear to be less detailed and therefore less useful, than corpus-based models derived from single characters. A second, perceptual experiment examines the relationship between the corpus-based models and personality perceptions of the original characters. We then summarize the paper and discuss future work.

Narrative and Dialogue Representation

In addition to models for ENLG that express differences in dramatic character, we also require a narrative story representation that can support an NLG engine. The standard architecture for NLG assumes a deep representation of meaning underlying utterances so that the input to an NLG engine typically consists of either concepts, speech acts, or communicative goals. An NLG engine is composed of the following modules and functions (Reiter and Dale 2000):

1. Content planning: refine communicative goals, select and structure content;

Speech Act:	ACCUSE
Relations:	JUSTIFY (nuc:1, sat:2); JUSTIFY (nuc:1, sat:3); JUSTIFY (nuc:1, sat:4);
Content:	1. assert(murderer (*Otter*))
	2. assert(is (*Otter*, impulsive))
	3. assert(stole (*Otter*, clams-c1))
	4. assert(not (like *Otter*, *Tortoise*))

Figure 5: A context plan for a scene in *Murder*, showing the speech act *accuse*, the discourse relation *justify* between content items, and the potential content pool, the assertions.

2. Sentence planning: choose linguistic resources (lexicon, syntax) to achieve goals;

3. Realisation: use grammar (syntax, morphology) to generate surface utterances.

In previous work, we argued that the story and its dialogue interactions should be represented as sequences of interleaved domain actions and speech acts. These representations can either be authored (Walker, Cahn, and Whittaker 1997a), or generated automatically via a planning mechanism (Riedl and Young 2004; André et al. 2000). This level of representation defines dependencies and discourse relations between plot elements. A second requirement is that each proposition in this representation have at least one elementary syntactic representation entered into a generation dictionary, and these syntactic representations need to be semantically indexed in such a way as to support different bases for content selection. For example, content can be indexed by positive and negative polarity, by who knows it or who is allowed to know it, among other narrative indices.

For each utterance then, this representation provides a high level communicative intention (speech-act) to be achieved, and a content pool that the ENLG can select from to achieve it. Then variation can be produced by either varying the parameters of the content selection mechanism, or those that vary the form of the selected content. For example, Fig. 3 provides a domain-independent, plan-based representation of a REQUEST-ACT (i.e. Laszlo's request for two cointreaux) (Litman and Allen 1990). The domain-specific content is represented by the 'action' variable in the definition of REQUEST-ACT (See Fig. 3), or a 'proposition' variable for an INFORM speech act. Thus, specific domains are represented in terms of the actions and propositions of that domain, as commonly used by planning representations. Fig. 4 illustrates the domain plan for serving two cointreaux, from the *Casablanca* domain. The ENLG operates on both the speech-act and the domain action representations, as long as there are dictionary entries for both levels of representation.

We are currently building a tool for constructing generation dictionaries drawing on the work of (Elson and McKeown 2007; 2009), which uses VerbNet and WordNet to anchor each word to its word sense in those semantic word hierarchies. We hope that our tool, like Scheherazade, will allow authors who do not have deep linguistic knowledge to semi-automatically construct dictionaries for their stories that will support a wide range of linguistic variation.

The story we are currently building, *Murder in the Arboretum*, is an outdoor augmented reality game with a fairly simple narrative structure of a murder mystery. The player is the detective, who must go from place to place to interview suspects and examine evidence. The game setting is the UCSC arboretum, and locations in the game are mapped onto locations in the arboretum. Game characters are derived from our previous work on SpyFeet (Reed et al. 2011). The input to the NLG engine is a context plan. A context plan representation for an *accuse* speech-act is shown in Fig. 5.

Two Approaches to Linguistic Style

Big Five Theory of Personality. A primary motivation for the Big Five model is that personality trait descriptions are pervasive in descriptions of dramatic and literary character (Allport 1960):

> Almost all the literature of character—whether [nonfiction] or fiction, drama or biography—proceeds on the psychological assumption that each character has certain *traits* peculiar to himself which can be defined through the narrating of typical episodes from life.

Another of Allport's observations was that traits important for describing differences in human behavior will have a corresponding lexical token, which is typically an adjective, e.g. *trustworthy, modest, friendly, spontaneous, talkative, dutiful, anxious, impulsive, vulnerable*. Allport and Odbert (1936) collected 17,953 trait terms from English and identified 4,500 as stable traits. Subsequent work analyzed how traits factor together in descriptions of people, leading to a standard framework of the Big Five personality traits as a way to describe essential personality differences among humans (Norman 1963; Goldberg 1990). The Big Five traits and some prototypical adjectives representing each end of each trait scale:

- **Extroversion**: warm, gregarious, assertive vs. shy, passive, joyless;
- **Emotional stability**: calm, even-tempered, reliable vs. neurotic, self-conscious, oversensitive;
- **Agreeableness**: trustworthy, considerate, friendly vs. selfish, suspicious, uncooperative;
- **Conscientiousness**: competent, disciplined, dutiful vs. disorganized, impulsive, unreliable;
- **Openness to experience**: intellectual, imaginative, curious vs. narrow-minded, ignorant, simple.

The advantages of the Big Five are that:

- It provides a concise framework for building models to control the linguistic style of narrative characters (5 main parameters, one for each trait);
- There are a number of validated questionnaires that can be used to evaluate human perceptions of character utterances;
- Psychologists have established the relation between the Big Five and other dimensions of expression variation, such as emotion. For example, there are strong relations between the extroversion and conscientiousness traits and the positive affects, and between neuroticism and disagreeableness and various negative affects (Watson and Clark 1992);
- Psychologists have documented a number of behavioral markers associated with each dimension, involving many aspects of communication such as language, speech, gesture and facial display. For example, previous research shows that extroverts talk more, louder, with more repetitions, positive words, faster movements, a firmer handshake and a more attractive smile, whereas neurotics produce more self-references, self-touching, disfluent gestures, negative emotion words and filled pauses (Furnham 1990; Pennebaker and King 1999; Scherer 1979; Doucet and Stelmack 1997; Dong et al. 1999; Chaplin et al. 2000; Gill and Oberlander 2003).

In addition, previous work has tested PERSONAGE both on its own and combined with text-to-speech, and facial expressions gesture engines based on the Big Five and shown that, in the case of restaurant recommendations, generated utterances are perceived by humans as expressing the intended personality traits (Neff et al. 2010; 2011; Bee et al. 2010; Mairesse and Walker 2010; 2011). We build on that work below, and generate utterances expressing particular personalities for dialogue in the INS *Murder in the Arboretum*.

Corpus-Based Models from Film. The second approach is to examine how **authors** actually operationalize character when writing dialogue, through an automatic corpus-based analysis of film screenplays, such as the examples in Fig. 6 from Annie Hall and Indiana Jones. To our knowledge, we are the first to analyze theatrical or film dialogue using natural language processing to derive computational models of character (Oberlander and Brew 2000; Vogel and Lynch 2008; Ireland and Pennebaker 2011). The general idea is to learn models of character linguistic style by counting linguistic reflexes (features) in film dialogue, and then use these learned models to control the parameters of the PERSONAGE generator. The PERSONAGE generator and its parameters are described in detail elsewhere (Mairesse and Walker 2011; 2010), as is a detailed description of our corpus-based learning method (Lin and Walker 2011).

There are many different ways we could learn such models (Isard, Brockmann, and Oberlander 2006; Walker, Rambow, and Rogati 2002; Walker et al. 2007). Here, we estimate models using vectors of features representing individual characters, and then derive distinctive features for that character by normalizing these feature counts against a representative population. For each feature x_i, the normalized value z_i is calculated as:

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

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_{xi}. 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

ANNIE HALL SCENE: Lobby of Sports Club	INDIANA JONES SCENE: Marion's Bar on Fire
ALVY: Uh ... you-you wanna lift? **ANNIE:** *Turning and aiming her thumb over her shoulder* Oh, why-uh ... y-y-you gotta car? **ALVY:** No, um ... I was gonna take a cab. **ANNIE:** *Laughing* Oh, no, I have a car. **ALVY:** You have a car? *Annie smiles, hands folded in front of her* **ALVY:** So ... *Clears his throat*. **ALVY:** I don't understand why ... if you have a car, so then-then wh-why did you say "Do you have a car?"... like you wanted a lift?	**INDY:** Let's get out of here! **MARION:** Not without that piece you want! **INDY:** It's here? *Marion nods, kicks aside a burning chair. Another burning beam falls from the roof. Indy pulls Marion close to him protectively.* **INDY:** Forget it! I want you out of here. Now! *He begins dragging her out.* **MARION:** *pointing*. There! *She breaks away from him, darts back and picks the hot medallion up in the loose cloth of her blouse.* **INDY:** Let's go! **MARION:** (looking around) You burned down my place! **INDY:** I owe you plenty! **MARION:** You owe me plenty! **INDY:** *smiles* You're something! **MARION:** I am something. And I'll tell you exactly what - *She holds up the medallion possessively.* I'm your partner!

Figure 6: Scenes from *Annie Hall* and *Indiana Jones and the Raiders of the Lost Ark*.

all of the characters of the same gender. Any Z-score >1 or <-1 is more than one standard deviation away from the mean. Z-scores greater 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 >1 or <-1 into one or more PERSONAGE generation parameters.

One advantage of this approach is that it lets us indirectly incorporate observations about types of characters from Archetype Theory. Archetype Theory provides a number of stock characters, such as HERO, SHADOW, or CAREGIVER, who have typical roles and personalities that can be re-used in different types of narrative. Rowe, Ha, and Lester (2008) produce heuristic models of character behavior using a taxonomy of 45 Master Archetypes (Rowe, Ha, and Lester 2008; Schmidt 2007), and show how archetype models can be integrated with dialogue models. However, when attempting to build on this approach, our perception was that taxonomies of character archetypes are difficult to operationalize; this is not surprising since their primary aim is to assist the writing practice of authors, rather than to offer a detailed inventory of parameters and models to control them in a computational framework.

We carried out an annotation study on a number of characters and scenes in our IMSDb (Internet Movie Script Database) corpus. The idea was to first classify film characters into particular archetypes, and then derive corpus-based models from the archetypes. We asked 3 annotators to classify 17 film characters into one of the 13 archetypes described in (Faber and Mayer 2009). The list of film characters and archetypes are in Table 1.

Our results from this pilot annotation project were not promising. For 5 out of 17 characters, there was no agreement among annotators: Costello (Ruler, Shadow, Everyman/Everywoman), Tyler (Hero, Magician, Jester), Agnis (Caregiver, Everyman/Everywoman, Sage), Goose (Hero, Jester, Explorer), and Spud (Everyman/Everywoman, Jester, Creator). The remaining 12 characters have at least 2 out of 3 in agreement. All together we achieved 53% agreement, with Cohen's kappa coefficient of 0.2921. This is very poor agreement. Thus, as an alternative test of this idea, we hand-selected several film characters that exemplify the "action hero" archetype with 27 male leads (Table 2). We hypothesized that all of these characters would instantiate the HERO archetype. Then we examine the effect of combining characters according to their archetype, and test whether the models derived for the archetype appear to be better when based on a set of characters than a single character.

Film Characters (17): Bruce: *Batman Returns*, Rae: *Black Snake Moan*, Neil: *Dead Poets Society*, Costello: *The Departed*, Tyler: *Fight Club*, Carter: *Final Destination*, Hooper: *Jaws*, Scott Smith: *Milk*, Furious: *Mystery Men*, Pete: *O Brother, Where Art Thou?*, Morris: *Purple Rain*, Paul: *Rachel Getting Married*, Plato: *Rebel without a cause*, Agnis: *The Shipping News*, Rose: *Titanic*, Goose: *Top Gun*, Spud: *Transpotting*
Archetypes (13): Caregiver, Creator, Everyman/Everywoman, Explorer, Hero, Innocent, Jester, Lover, Magician, Outlaw, Ruler, Sage, Shadow

Table 1: Annotation Task Film Characters and Archetypes

McClane: *Die Hard*, Rambo: *Rambo*, Peter and Spider-Man: *Spider-Man*, Bourne: *The Bourne* series (three movies), Indiana Jones: *Indiana Jones* series (first three movies), Maximus: *Gladiator*, Han and Luke: *Star Wars: Episode VI - Return of the Jedi* and *Star Wars: Episode V - The Empire Strikes Back*, Anakin and Obi-Wan: *Star Wars: Episode II - Attack of the Clones*, Luke and Han: *Star Wars*, Rafe and Danny: *Pearl Harbor*, Burnett and Lowrey: *Bad Boys*, and Goodspeed and Mason: *The Rock*.

Table 2: 27 Hand Selected Action Heroes

Deriving and Evaluating Character Models

Our goal is to be able to compare different types of models of character. Here our focus is to compare corpus-based and personality-based models of character. We hope to develop a series of objective metrics that will be predictive of the quality of a model, obviating the need to do detailed perceptual experiments on each model. However at the moment,

Segment	Ace	Indy	Jack	Carrie	Hermione	Jackie
1 (∼100 turns)	33	33	34	38	36	30
2 (∼200 turns)	32	34	35	31	36	34
3 (∼300 turns)	36	35	36	35	36	34
4 (∼400 turns)	34	33	36	33	36	32
5 (∼500 turns)	34	34	35	35	–	–
6 (∼600 turns)	34	34	36	–	–	–
7 (∼700 turns)	35	36	–	–	–	–

Table 3: Number of Significant Attributes based on Dialogue Turns for z>1 and z<-1

our primary objective metric for model qualityl, apart from perceptual experiments, is the number of parameters in the model that indicate significant differences in linguistic style, altogether, and at each level of statistical significance.

As mentioned above, there are different ways of deriving corpus-based models, e.g. characters can be grouped by their archetype, they can be grouped by their film genre or by gender, or we can derive a model for a single character. We first wished to explore models derived by selecting a single character and normalizing their linguistic behavior against all the characters of the same gender. Then we wished to examine the effect of corpus size on the specificity of models.

It would be expected that the more dialogue turns are used for the model, the better the model would be. Thus we first examine the effect of number of dialogue turns on the number of significant attributes in the models. We look at male characters Ace from *Casino* with 747 turns, Indiana Jones from the three *Indiana* series with 776 turns, and Jack from *Fight Club* with 626 turns. For female characters, we chose Carrie from *Sex and the City* with 518 turns, Hermione from the *Harry Potter* series with 481 turns, and Jackie from *Stepmom* with 444 turns. They were chosen based on the large number of their dialogue turns compared to other characters.

The turns were randomized and separated into incrementing segments of roughly 100 turns. For example, Indiana has 776 turns, separated into 7 segments where segment 1 contains the first 110 turns, segment 2 contains the first 110 turns plus the next 111 turns, etc. Segment 7 contains all 776 turns.

We found that the number of significant attributes >1 and <-1 stayed relatively the same regardless of the number of turns for all characters, as shown in Table 3. However, as we increase the cutoff to z>3 and z<-3, there is a trend that the more turns a character has, the more significant attributes there are in the resulting learned model. Fig. 7 shows trends for male and female characters, as well as individual characters Indiana Jones and Carrie from *Sex and the City*. From these individual characters' plots, we can see that z>2 and z<-2, as well as z>3 and z<-3, show an upward trend.

Archetype Experiment. The results suggest that combining characters by archetype does not improve our corpus-based models. The effect of number of turns/utterances does not hold when we combine these characters. Using combined utterances actually reduces the number of significant attributes than using a single character on its own. This suggests that these actions heroes might contain other styles of dialogue such comedy or drama, so that as a result of combining all these different characters, we see the resulting dialogue being "blended" within the whole male population. Table 4 show some of the character combinations, along with their significant attributes and Z-scores. LIWC (Linguistic Inquiry and Word Count) is a text analysis software program that counts categorical words. For example, LIWC-Cause counts causation words such as *because, effect, hence*.

Bourne: (10) LIWC Cause (z=2.05), LIWC Self (z=1.91), LIWC I (1.87), word *because* (1.48), LIWC Pronoun (1.44), LIWC Discrep (1.07), LIWC Sixltr (-1.09), LIWC WPS (-1.16), LIWC Posemo (-1.21), word *though* (-1.41)
Bourne + The Rock: (9) word *since* (z=1.56), LIWC Period (z=1.31), LIWC I (z=1.13), LIWC Self (z=1.09), word *because* (z=1.02), LIWC WPS (z=-1.01), Reject first (z=-1.02), word *though* (z=-1.41), word *so* (z=-1.51)
Bourne + The Rock + Independence Day: (7) word *because* (z=1.29), word *since* (z=1.24), LIWC Period (z=1.13), LIWC Posemo (z=-1.01), word *but* (z=-1.30), word *though* (z=-1.41), word *so* (z=-1.53)
Bourne + The Rock + Independence Day + Die Hard: (5) word *because* (z=1.12) word *since* (z=1.07) word *but* (z=-1.10) word *so* (z=-1.31) word *though* (z=-1.41)
All: (2) phrase *it seems to me* (z=1.04), word "though" (z=-1.20)

Table 4: Action Heroes Significant Attributes and Z-scores

We also compared models that result from normalizing a specific action hero against a particular genre population. We noticed that this type of normalization brings out more specific and detailed model parameters than when we normalize against all males. For example, Bourne compared to all males resulted in 10 significant attributes (usage of *because* and *though*, LIWC's Cause, Self, I, Pronoun, Discrep, Sixltr, WPS, and Posemo), but when compared to comedy males, there were 24 significant attributes (e.g., LIWC's Negate, Assent, Qmarks, Unique, Preps).

Table 5 illustrates the result of applying example character models to *Murder* utterances, and some of the variations in style that we are currently able to produce. The refinement of the mapping and development of the generation dictionary for *Murder* is work in progress.

Perceptual Experiment

We also carried out a perceptual experiment to directly compare the perceptions of personality for characters represented by 3 scenes from films, illustrating the utterance styles of 6 characters (3 male and 3 female). We collect user perceptions of the personality of those characters using the Ten Item Personality Inventory (TIPI) (Gosling, Rentfrow, and Swann 2003). The TIPI returns values between 1 to 7 for each trait, with each trait having a low end (introversion) versus high end (extroversion).

Using the PERSONAGE generator, we generate dialogic utterances for the characters in the story of *Murder*, using both (1) the film character model; and (2) six rule-based personality models from previous work (high and low values for extroversion, emotional stability and agreeableness) (Mairesse and Walker 2010). We collect user perceptions of the personality of *Murder* characters whose linguistic style is controlled by these models (6 film character models and 6 personality models), again using the TIPI. We predict that:

- H1: The rule-based models for personality expression (previously tested only in the restaurant recommendation

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

Table 5: Utterances for *Murder* generated using Film Character Models

domain), will be perceived as expressing that personality in our story domain.

- H2: Utterances generated using character models will correlate more strongly with character personality perceptions than utterances generated using rule-based models for personality expression.

The 6 film characters used for our study are Alvy and Annie (*Annie Hall*), Indy and Marion (*Indiana Jones - Raiders of the Lost Ark*), and Vincent and Mia (*Pulp Fiction*).

Results. 29 subjects (13 female and 16 male, ages ranging from 22 to 44) participated in a web-based experiment. Table 6 shows the mean values of the TIPI scale judgments for Big Five traits of Extroversion, Emotional Stability and Agreeableness for the 6 characters judged on original utterances and generated utterances.

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 determine whether characters were perceived as having distinct personalities (within subjects). Using 5% as the significance threshold, the results shown in Table 8 indicate that the personality of Alvy is perceived as significantly different from all of the other characters, and Annie is perceived as signif-

Trait	Alvy	Annie	Indy	Marion	Mia	Vincent
	Original Utterances					
Extroversion	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
	Corpus-Based Models					
Extroversion	3.3	3.4	4.8	4.8	3.3	3.7
Agreeableness	4.8	4.2	3.8	3.4	4.1	3.5
Emotional Stability	5.4	3.8	3.5	3.6	2.5	2.9

Table 6: Average Big Five Scores for Original Utterances and Corpus-Based Models on TIPI Scale of 1 to 7

Trait	High	Low	P-value
Extroversion	5.2	3.3	<0.001
Emotional Stability	5.5	2.7	<0.001
Agreeableness	3.4	3.4	–

Table 7: Big Five Personality Scores for *Murder* Utterances generated using Big Five Models

Char	Marion	Vincent	Mia	Alvy	Annie
	Original Utterances				
Indy	0.196	0.578	0.138	**<0.001**	0.163
Marion	–	0.463	0.814	**0.002**	**0.002**
Vincent	–	–	0.403	**<0.001**	**0.026**
Mia	–	–	–	**<0.001**	**0.003**
Alvy	–	–	–	–	**<0.001**
	Corpus-Based Models				
Indy	**<0.001**	**0.033**	**0.008**	**<0.001**	**<0.001**
Marion	–	0.232	0.522	**0.002**	**0.024**
Vincent	–	–	0.485	**<0.001**	**<0.001**
Mia	–	–	–	**0.001**	**0.001**
Alvy	–	–	–	–	0.548

Table 8: P-values for Comparing the Similarities of the Perceived Personality of Characters Based on Original Utterances and Based on the Corpus-Based Models

icantly different from all characters except for Indy. These results suggest that the differences in perceived personality across different characters are small, with the Tarantino (Vincent, Mia) and Spielberg (Indy, Marion) characters being perceived as having similar personalities.

Our results for H1 are mixed. Table 7 shows the high and low values, as well as the paired t-test results comparing the high/low values of each trait. Both extroversion and emotional stability show significant differences ($p<0.001$). However, differences in high/low agreeableness were not perceived in the *Murder* domain. There are several possible reasons for this; perhaps the limited set of utterances tested do not show the variability in agreeableness that the PERSONAGE generator is capable of, or perhaps manifesting agreeableness in the *Murder* domain requires the addition of new parameters to the PERSONAGE generator.

In a similar calculation, using the corpus-based models in Table 8, we see that both Alvy and Indy are perceived as significantly different from other characters, and Annie is perceived as significantly different from Indy rather than Alvy. This might imply that our corpus-based models extracted

Corpus-Based	Original Utterances					
	Indy	Marion	Vincent	Mia	Alvy	Annie
Indy	**0.14***	-0.30	-0.08	-0.12	-0.14	-0.32
Marian	-0.03	**-0.07**	-0.10	-0.11	0.06	0.07
Vincent	-0.18	0.06	**0.36***	0.05	0.12	0.25
Mia	-0.05	0.44*	0.26	**0.19***	0.04	0.31
Alvy	-0.30	0.01	-0.05	-0.11	**0.45***	0.45*
Annie	-0.22	0.25	-0.01	0.10	0.32	**0.35**

*=strongest positive correlation

Table 9: Correlations between Personality Judgments Original Utterances and Corpus-Based Models

and emphasized only certain features of the film characters.

Next, we compare perceived personality of original and corpus-based utterances. Again with combined traits, the correlation is shown in Table 9. Focusing on the positive correlation, the results show that the same film character from corpus-based and original utterances often correspond more highly to each other than to other characters. Alvy has the strongest self-correlation (0.45), followed by Vincent (0.36), Mia (0.19), and Indy (0.14). However, Annie's model corresponds more strongly to Alby (0.35), while Marion's model is negative correlated with her own character and correlating strongly to Mia (0.44). This shows that our corpus-based generated utterances correspond more to the original utterances of the same character than to different characters.

H2 is only weakly supported. The correlations and P-values between the corpus-based generated and original utterances are shown in Table 10. The correlation coefficients show that Alvy's extroversion and emotional stability, Annie's emotional stability, and Vincent's agreeableness have the highest positive correlations for corpus-based models. All other characters and traits correlate more strongly in the rule-based models.

However, we believe that the current results are encouraging. For models with good confidence in correlation and/or significant P-values, the corpus-based models are more specific and detailed than rule-based models.

Discussion

This paper has examined two different sources for models of character linguistic style to use for character dialogue in interactive narrative. We have explored the use of these models in the context of our pilot outdoor augmented reality game *Murder in the Arboretum*. We show that corpus-based character models based on a single character are more detailed and specific than either personality based models or models based on a collection of characters exemplifying the same dramatic archetype, at least with respect to the way we have compared these different sources of models.

In future work, we intend to explore a more detailed analysis of our corpus of screenplays, for example to look at the effect of different types of contextual variables on a character's linguistic style (Elson and McKeown 2010). We would like to model some aspects of a character's relationships and see how that affects their dialogue interaction. Politeness theory would predict that relationship of the conversant will have a major effect on linguistic style (Gupta, Walker, and Romano 2007).

While we have not used politeness theory here, it would be useful to incorporate some of its observations. One of its advantages is that it explicitly models aspects of the relationship between speakers, and it represents the causal links between indirect speech acts and their direct counterparts. However it has several limitations in our view: (1) the models for controlling the parameters are only based on social distance and power; and (2) the parameters (e.g. use of hedges, indirect speech acts) are defined at a somewhat course level; (3) theories of politeness do not necessarily map onto the way that authors of interactive stories think about character or dialogue. We would also like to further explore blended models, for example overlays of character-based corpus models and rule-based personality models.

References

Allport, G. W., and Odbert, H. S. 1936. Trait names: a psycho-lexical study. *Psychological Monographs* 47(1, Whole No. 211):171–220.

Allport, G. W. 1960. *Personality and social encounter*. Boston, MA: Beacon.

André, E.; Rist, T.; van Mulken, S.; Klesen, M.; and Baldes, S. 2000. The automated design of believable dialogues for animated presentation teams. In *Embodied conversational agents*. Cambridge, MA: MIT Press. 220–255.

Bee, N.; Pollock, C.; André, E.; and Walker, M. 2010. Bossy or wimpy: expressing social dominance by combining gaze and linguistic behaviors. In *Intelligent Virtual Agents*, 265–271. Springer.

Brennan, S., and Ohaeri, J. 1994. Effects of message style on users' attributions toward agents. *Conf. on Human Factors in Computing Systems* 281–282.

Brown, P., and Levinson, S. 1987. *Politeness: Some universals in language usage*. Cambridge University Press.

Cassell, J., and Bickmore, T. 2003. Negotiated collusion: Modeling social language and its relationship effects in intelligent agents. *User Modeling and User-Adapted Interaction* 13:89–132.

Cavazza, M., and Charles, F. 2005. Dialogue generation in character-based interactive storytelling. In *AAAI First Annual AIIDE Conf.*.

Chaplin, W. F.; Phillips, J. B.; Brown, J. D.; Clanton, N. R.; and Stein, J. L. 2000. Handshaking, gender, personality, and first impressions. *Journal of Personality and Social Psychology* 79(1):110–117.

Dong, J. K.; Jin, T. H.; Cho, H. W.; and Oh, S. C. 1999. The esthetics of the smile: a review of some recent studies. *Int. Journal of Prosthodontics* 12(1):9–19.

Doucet, C., and Stelmack, R. M. 1997. Movement time differentiates extraverts from introverts. *Personality and Individual Differences* 23(5):775–786.

Elson, D., and McKeown, K. 2007. A platform for symbolically encoding human narratives. In *Proc. of the AAAI Fall Symposium on Intelligent Narrative Technologies*.

	Actual Film Utterances																	
	Indy			Marion			Vincent			Mia			Alvy			Annie		
	EXT	AGR	EMS	EXT	AGR	EMS	EXT	AGR	EMS	EXT	AGR	EMS	EXT	AGR	EMS	EXT	AGR	EMS
Corpus	0.05†	-0.13‡	-0.10	0.11‡	0.03	0.07	0.21	0.36	0.14	0.10	-0.06†	0.13†	0.38	-0.24	0.39†	-0.03	0.23‡	0.49
Rule High	0.04‡	-0.12	0.27†	-0.05	0.12	0.60‡	-0.24	0.25	0.08‡	0.14	-0.02	0.21‡	-0.27‡	-0.07	-0.06‡	0.20†	0.05‡	0.16‡
Rule Low	0.20†	0.20	-0.45‡	0.14	0.43†	0.07‡	0.28‡	0.34†	-0.14‡	0.51‡	0.09	0.05‡	0.24	-0.08	0.01‡	0.32‡	0.42‡	-0.05

EXT=extroversion, AGR=agreeableness, EMS=emotional stability; †= p<0.05, ‡= p<0.01

Table 10: Correlation between Corpus-Based and Rule-Based Models

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

Elson, D., and McKeown, K. 2010. Automatic attribution of quoted speech in literary narrative. In *Proc. of AAAI*.

Faber, M., and Mayer, J. 2009. Resonance to archetypes in media: There's some accounting for taste. In *Journal of Research in Personality*, volume 43, number 3. 307–322.

Furnham, A. 1990. Language and personality. In Giles, H., and Robinson, W., eds., *Handbook of Language and Social Psychology*. Winley.

Gill, A., and Oberlander, J. 2003. Perception of e-mail personality at zero-acquaintance: Extraversion takes care of itself; neuroticism is a worry. In *Proc. of the 25th Annual Conf. of the Cognitive Science Society*, 456–461.

Goldberg, L. R. 1990. An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology* 59:1216–1229.

Gosling, S. D.; Rentfrow, P. J.; and Swann, W. B. 2003. A very brief measure of the big five personality domains. *Journal of Research in Personality* 37:504–528.

Gupta, S.; Walker, M. A.; and Romano, D. M. 2007. How rude are you?: Evaluating politeness and affect in interaction. In *Proc. of ACHI*, 203–217.

Hayes-Roth, B., and Brownston, L. 1994. Multiagent collaboration in directed improvisation. Technical Report KSL 94-69, Knowledge Systems Laboratory, Stanford University.

Ireland, M., and Pennebaker, J. 2011. Authors' gender predicts their characters' language. *In submission*.

Isard, A.; Brockmann, C.; and Oberlander, J. 2006. Individuality and alignment in generated dialogues. In *Proc. of the 4th Int. Natural Language Generation Conf. (INLG-06)*, 22–29.

Isbister, K., and Nass, C. 2000. Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics. *Int. Journal of Human-Computer Studies* 53(2):251 – 267.

Johnson, L.; Mayer, R.; André, E.; and Rehm, M. 2005. Cross-cultural evaluation of politeness in tactics for pedagogical agents. In *Proc. of the 12th Int. Conf. on Artificial Intelligence in Education (AIED)*.

Lin, G. I., and Walker, M. A. 2011. All the world's a stage: Learning character models from film. In *Proc. of the Seventh AIIDE Conf.*.

Litman, D., and Allen, J. 1990. Recognizing and relating discourse intentions and task-oriented plans. In Cohen, P.; Morgan, J.; and Pollack, M., eds., *Intentions in Communication*. MIT Press.

Louchart, S.; Aylett, R.; Dias, J.; and Paiva, A. 2005. Unscripted Narrative for affectively driven characters. *Proc. First Int. Conf. on AIIDE*.

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

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

Mateas, M., and Stern, A. 2003. Façade: An experiment in building a fully-realized interactive drama. In *Proc. of the Game Developers Conf., Game Design track*.

Mateas, M. 2007. The authoring bottleneck in creating AI-based interactive stories. In *Proc. of the AAAI 2007 Fall Symposium on Intelligent Narrative Technologies*.

Mott, B., and Lester, J. 2006. U-director: a decision-theoretic narrative planning architecture for storytelling environments. *Proc. of the Fifth Int. Joint Conf. on Autonomous agents and multiagent systems* 977–984.

Murray, J. 1997. *Hamlet on the Holodeck: The Future of Narrative in Cyberspace*. The Free Press New York, NY, USA.

Neff, M.; Wang, Y.; Abbott, R.; and Walker, M. 2010. Evaluating the effect of gesture and language on personality perception in conversational agents. In *Intelligent Virtual Agents*, 222–235. Springer.

Neff, M.; Toothman, N.; Bowmani, R.; Fox Tree, J. E.; and Walker, M. 2011. Don't Scratch! Self-adaptors Reflect Emotional Stability. In *Proc. of Intelligent Virtual Agents*. Springer.

Norman, W. T. 1963. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality rating. *Journal of Abnormal and Social Psychology* 66:574–583.

Oberlander, J., and Brew, C. 2000. Stochastic text generation. *Philosophical Transactions of the Royal Society of London* Series A, 358:1373–1385.

Pennebaker, J. W., and King, L. A. 1999. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology* 77:1296–1312.

Piwek, P. 2003. A flexible pragmatics-driven language generator for animated agents. In *Proc. of EACL*.

Rapp, D.; Gerrig, R.; and Prentice, D. 2001. Readers' Trait-Based Models of Characters in Narrative Comprehension. *Journal of Memory and Language* 45(4):737–750.

Reed, A.; Samuel, B.; Sullivan, A.; Grant, R.; Grow, A.; Lazaro, J.; Mahal, J.; Kurniawan, S.; Walker, M.; and Wardrip-Fruin, N. 2011. A step towards the future of role-playing games: The spyfeet mobile rpg project. In *Proc. of the Seventh AIIDE Conf.*. AAAI.

Reiter, E., and Dale, R. 2000. *Building Natural Language Generation Systems*. Cambridge University Press.

Riedl, M., and Young, R. M. 2004. An intent-driven planner for multi-agent story generation. In *Proc. of the 3rd Int. Conf. on Autonomous Agents and Multi Agent Systems*.

Ross, L. 1977. The intuitive psychologist and his shortcomings: Distortions in the attribution process. *Advances in experimental social psychology* 10:173–220.

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

Scherer, K. R. 1979. Personality markers in speech. In Scherer, K. R., and Giles, H., eds., *Social markers in speech*. Cambridge University Press. 147–209.

Schmidt, V. 2007. *45 Master characters*. Writers Digest Books.

Short, E. 2009. Opinion: Heileen and the art of game storytelling.

Thue, D.; Bulitko, V.; and Spetch, M. 2008. Making stories player-specific: Delayed authoring in interactive storytelling. In *Interactive Storytelling*, volume 5334 of *Lecture Notes in Computer Science*. 230–241.

Vogel, C., and Lynch, G. 2008. Computational Stylometry: Whos in a Play? *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction* 169–186.

Walker, M. A., and Rambow, O. 2002. Spoken language generation. *Computer Speech and Language, Special Issue on Spoken Language Generation* 16(3-4):273–281.

Walker, M. A.; Stent, A.; Mairesse, F.; and Prasad, R. 2007. Individual and domain adaptation in sentence planning for dialogue. *Journal of Artificial Intelligence Research (JAIR)* 30:413–456.

Walker, M.; Cahn, J. E.; and Whittaker, S. J. 1997a. Improvising linguistic style: Social and affective bases for agent personality. In *Proc. of the 1st Conf. on Automous Agents*, 96–105.

Walker, M.; Rambow, O.; and Rogati, M. 2002. Training a sentence planner for spoken dialogue using boosting. *Computer Speech and Language* 16(3-4).

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

Watson, D., and Clark, L. A. 1992. On traits and temperament: General and specific factors of emotional experience and their relation to the five factor model. *Journal of Personality* 60(2):441–76.

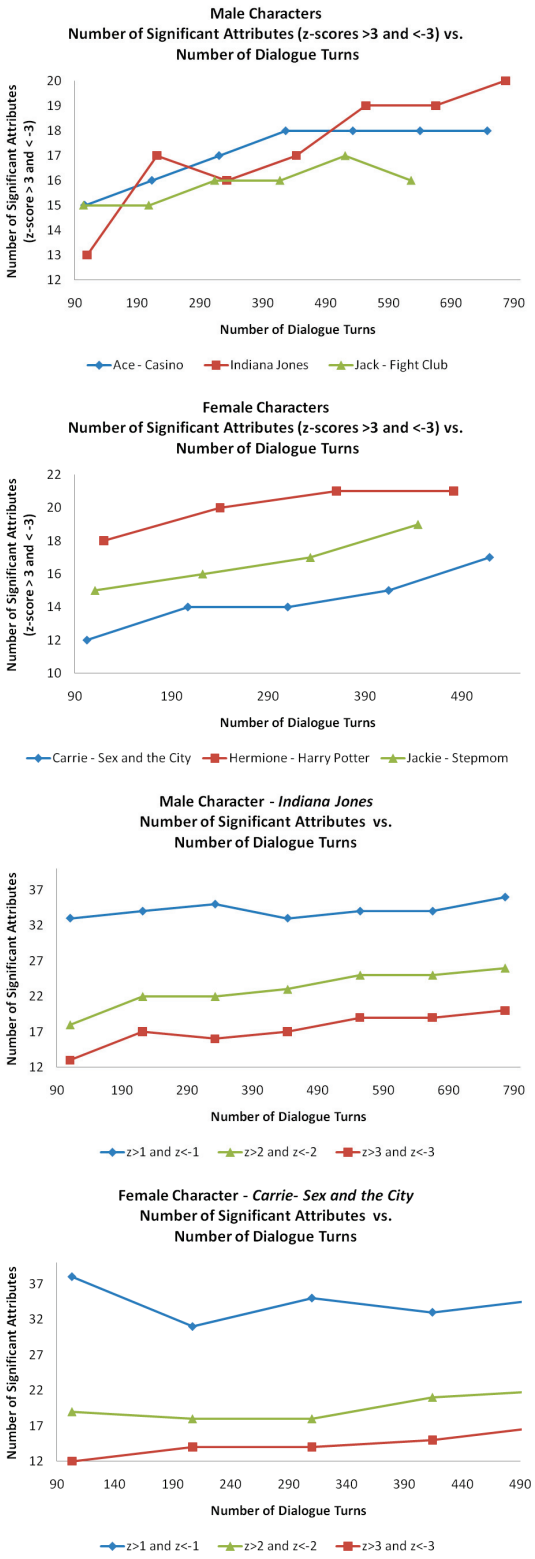

Figure 7: Trends for Number of Significant Attributes and Dialogue Turns for Male and Female Characters